2013

Expectation-maximization algorithms for learning a finite mixture of univariate survival time distributions from partially specified class values

Youngrok Lee
Iowa State University

Expectation-maximization algorithms for learning a finite mixture of univariate survival time distributions from partially specified class values

by

Youngrok Lee

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:

Sigurdur Olafsson, Major Professor

John Jackman

Lizhi Wang

Dianne Cook

Dan Zhu

Iowa State University

Ames, Iowa

2013

DEDICATION

I would like to dedicate this thesis to my wife Min-Jeong Yang without whose support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First I would like to heartily express the deepest appreciation to Dr. Sigurdur Olafsson for his persistent guidance, patience and support throughout this dissertation research. His insights and words of encouragement have helped me keep concentrating on this work whenever I was loosing confidence. I would also like to thank my committee members for their efforts and contributions to this work: Dr. John Jackman, Dr. Lizhi Wang, Dr. Dianne Cook and Dr. Dan Zhu. I am also thankful to Dr. Jae-Kwang Kim for his helpful advices improved this research.

I would also like to greatly thank to the Midwest Forensics Resource Center at Ames Laboratory for its financial support during my study. This research project was funded by the U.S. Department of Justice, National Institute of Justice, through the Midwest Forensics Resource Center at Ames Laboratory, under Cooperative Agreement number 2009-D1-BX-K206. The Ames Laboratory is operated for the U.S. Department of Energy by Iowa State University, under contract No. DE-ACO2-07CHI 1 358.

ABSTRACT

Heterogeneity exists on a data set when samples from different classes are merged into the data set. Finite mixture models can be used to represent a survival time distribution on heterogeneous patient group by the proportions of each class and by the survival time distribution within each class as well. The heterogeneous data set cannot be explicitly decomposed to homogeneous subgroups unless all the samples are precisely labeled by their origin classes; such impossibility of decomposition is a barrier to overcome for estimating finite mixture models. The expectation-maximization (EM) algorithm has been used to obtain maximum likelihood estimates of finite mixture models by soft-decomposition of heterogeneous samples without labels for a subset or the entire set of data. In medical surveillance databases we can find partially labeled data, that is, while not completely unlabeled there is only imprecise information about class values. In this study we propose new EM algorithms that take advantages of using such partial labels, and thus incorporate more information than traditional EM algorithms. We particularly propose four variants of the EM algorithm named EM-OCML, EM-PCML, EM-HCML and EM-CPCML, each of which assumes a specific mechanism of missing class values. We conducted a simulation study on exponential survival trees with five classes and showed that the advantages of incorporating substantial amount of partially labeled data can be highly significant. We also showed model selection based on AIC values fairly works to select the best proposed algorithm on each specific data set. A case study on a real-world data set of gastric cancer provided by Surveillance, Epidemiology and End Results (SEER) program showed a superiority of EM-CPCML to not only the other proposed EM algorithms but also conventional supervised, unsupervised and semi-supervised learning algorithms.

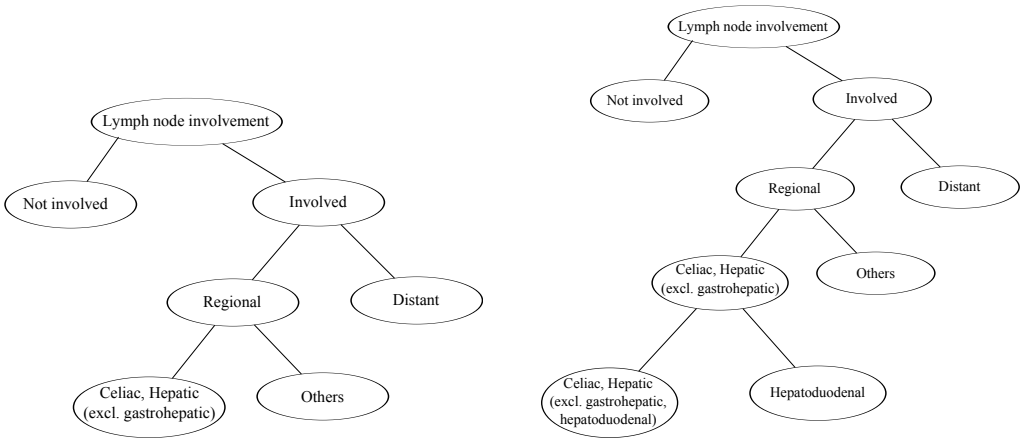
CHAPTER 1. INTRODUCTION

This chapter introduces research problems that are dealt with by this thesis as well as expected benefits of this study.

1.1 Motivation

"How long will I survive?" may be one of the most frequently asked question from cancer patients. Answers to such questions come from observed survival time of similarly diagnosed patents in the past. Surveillance, Epidemiology and End Results (SEER) program (www.seer.cancer.gov) of the National Cancer Institute provides tremendous amount cancer cases that have been collected since 1973 in the United States. So fully utilizing the SEER data is desirable to give cancer patients precise expectation about their remaining time to death. Diagnostic information like contiguous extension of primary tumor or involvement of lymph nodes is particularly important to be utilized because expected survival time significantly depends on such diagnostic information.

Medical surveillance data often fail to deliver precise diagnostic information. From 1988 to 2003, for example, every instance of the gastric signet ring cell carcinoma on the cardia in the SEER database was desired to have values of {`Not involved`}, {`Celiac, Hepatic (excl. gastrohepatic)`}, {`Other regional`}, or {`Distant`}, which are corresponding to the leaf nodes of Figure 1.1(a). However more than half of the cases that have been collected from 1988 to 2003 contain imprecise information about the lymph node involvement: {`Unknown`}, {`Involved, NOS`}, and {`Regional, NOS`} (Table 1.1). A change of the procedure of collecting data over years is an important characteristic of SEER database. Figure 1.1(b) represents a hierarchy of values for the lymph node involvement applied from 2004 to 2008; it further spec-



(a) 1988–2003 (b) 2004–2008

Figure 1.1: Categorization of values on the lymph node involvement attribute for the gastric signet ring cell carcinoma on the cardia in SEER research data. (a) Categorization applied from 1988 to 2003. (b) Categorization applied from 2004 to 2008.

ified {`Celiac, Hepatic (excl. gastrohepatic)`} into {`Hepatoduodenal`} and {`Celiac, Hepatic (excl. gastrohepatic and hepatoduodenal)`}. With the newly extended hierarchy of attribute values, 29 cases of {`Celiac, Hepatic (excl. gastrohepatic)`} that were collected until 2003 are considered having imprecise diagnostic information, although they were fully specified cases at the time of being collected.

Completeness of cancer information is one of goals that SEER program is trying to achieve; so cases with imprecise diagnostic information are undesirable. They may however be unavoidable because of practical difficulties in input data control, updates of data coding systems with imperfect migration of old data, and so on. However, cases with imprecise diagnostic information still carry more specific information than cases with no diagnosis; and utilizing this while estimating survival time distributions may hence reduce uncertainties of the results. This study has been motivated to fully utilize such imprecise diagnostic information in estimating survival time distributions and to see how such uses of imprecise diagnostic information contribute to population studies of cancer patients.

Table 1.1: Number of observed values on the lymph node involvement attribute for the gastric signet ring cell carcinoma on the cardia in SEER research data. (*NOS: not otherwise specified)

Values	1988–2003	2004–2008
Unknown	186	26
Known		
Not involved	106	86
Involved, NOS*	3	1
Regional, NOS*	124	20
Celiac, Hepatic (excl. gastrohepatic)	29	
Celiac, Hepatic (excl. gastrohepatic, hepatoduodenal)		16
Hepatoduodenal		1
Other regional	92	86
Distant, NOS*	56	
Total	596	236

1.2 Toward Partially Supervised Learning of Survival Time Models

Databases are composed of samples that have been observed from a population. When the population is composed of several subpopulations, an instance in databases is expected to have *primary data* that explain characteristics of the instance and a *label* which represents the origin, that is, the subpopulation from which the instance was observed. Instances in a database have traditionally been categorized into *labeled* data and *unlabeled* data according to whether the labels were observed. Over the past decades there has been great interest in learning statistical knowledge about primary data within each subpopulation that makes it distinguishable to the other subpopulations. Unsupervised learning (e.g. clustering) is typically used when all the available data are unlabeled. On the other hand, supervised learning (e.g. classification) is appropriate to use when all the data are labeled. Semi-supervised learning has extended both unsupervised learning and supervised learning by using both labeled and unlabeled data in a single learning procedure (Zhu and Goldberg, 2009). Therefore, semi-supervised learning utilizes more information to contribute to knowledge discovery than supervised and unsupervised learning.

In real-world data, observed labels may carry only partial information about the origins of samples. Numerous such examples exist in a variety of applications. Consider for example a

Table 1.2: Types of data that are fully utilized in each learning method

Type of learning	Type of data		
	labeled	unlabeled	partially labeled
Supervised	√		
Unsupervised		√	
Semi-supervised	√	√	
Partially supervised	√	√	√

screenshot from a movie where the characters in the scene are known based on the dialogue. Then we may not be able to identify the name associated with a face exactly, but we can choose small numbers of candidates (Cour et al., 2009, 2011). Another example is where a component that caused system failure cannot be exactly identified due to cost or time constraints, a subset of components that contains the true cause could still be identified (Usher and Hodgson, 1988; Usher and Guess, 1989; Guess et al., 1991). Finally, in student databases some student status can be simply categorized into undergraduate and graduate, whereas some others have more specified information such as whether the graduate student is in master program or in doctorate program (Zhang, 2005; Zhang et al., 2006). All three examples represent data with labels that carry partial but incomplete information about the origin subpopulation. Such labels have been called *partial labels* (Ambroise and Govaert, 2000; Cour et al., 2011), *ambiguous labels* (Cour et al., 2009), or *imprecise labels* (Vannoorenberghe and Smets, 2005; Côme et al., 2008, 2009). We call samples *partially labeled* data if the samples consist of primary data and partial labels.

None of supervised, unsupervised, and semi-supervised learning methods cannot fully utilize partially labeled data for data mining. *Partially supervised learning*, which represents learning from partially labeled data, intends to take advantages of fully utilizing partially labeled data (Table 1.2). In the same ways that using unlabeled data leads to more reliable statistical knowledge when only few labeled data are given (Zhu and Goldberg, 2009), with partially supervised learning we can expect to take advantage of partially labeled data when only few labeled and unlabeled data are given. In the area of reliability engineering, learning failure time models from partially labeled data have been studied since the 1980's (Usher and Hodgson,

1988; Usher and Guess, 1989; Guess et al., 1991; Lin and Guess, 1994; Ramon et al., 1995; Park, 2005; Flehinger et al., 1996, 1998). In the field of machine learning however only few studies have been done regarding partially supervised learning for face recognition (Cour et al., 2009, 2011) and finite mixture modeling (Ambroise and Govaert, 2000).

Proposed methods in this study fully utilize partially labeled data in survival time analysis. This study considers survival time of a cancer patient as primary data, while diagnostic information labels cancer patients to group patients at a same risk together. As described in Section 1.1 a considerable number of cases in medical surveillance database are partially labeled with diagnostic information. This study intends to use such imprecise diagnostic information in estimating expected time to death or survival rates for subpopulation at each risk level. In particular we focus on estimating a finite mixture of survival time distributions. Learning finite mixture models from partially labeled data has already been studied by Ambroise and Govaert (2000). However Ambroise and Govaert (2000) utilized partially labeled data under a strict assumption on label observing mechanisms, although the assumption was not explicitly stated in their study. We therefore propose more generalized learning methods that can be useful when underlying assumptions on Ambroise and Govaert (2000) are violated. In addition, we propose a learning method with a specific label observing mechanism that possibly exists in a hierarchical structure of diagnostic information.

1.3 Organization of Dissertation

The remainder of this thesis is organized as follows. Chapter 2 introduces background on finite mixture models, partially supervised learning, and survival time models. In chapter 3 new Expectation-Maximization (EM) algorithms are proposed to estimate a mixture of survival time models from partially labeled data. In Chapter 4 the new EM algorithms are specialized for a hierarchical structure of observable labels called attribute value taxonomy (AVT). Chapter 4 also proposes a new EM algorithm for a specific type of label observing mechanism on AVT. In Chapter 5 We conduct experiments on synthetic data that is generated from a survival tree introduced by Davis and Anderson (1989) and conduct a case study on SEER database. Finally conclusions are made in Chapter 6.

CHAPTER 2. BACKGROUND

This chapter provides background knowledge about three major components of this study: finite mixture models, partially supervised learning, and survival time models.

2.1 Finite Mixture Models (FMM)

Finite mixture modeling is a study of learning how a heterogeneous population is composed of predetermined numbers of subpopulations (McLachlan and Peel, 2000). Finite mixture models are popularly and widely used when knowing statistical patterns of primary data within each subpopulation, as well as mixing proportions, produces valuable knowledge that cannot be known from the marginal patterns. For example, finite mixture models extract signals from noisy spectroscopy data by considering signal and noise to be separate subpopulations (Kuss et al., 2002). Also, failure time of a system can be specifically estimated for each possible cause (Mendenhall and Hader, 1958; Papadapoulos and Padgett, 1986). In addition, finite mixture models have been widely applied to medical data analysis (Schlattmann, 2008).

A finite mixture model is usually defined as a weighted sum of simple parametric densities, each of which represents a density of interest within a homogeneous subpopulation. With K subpopulations the mixture of densities of multivariate data $\mathbf{x} = (x_1, \cdots, x_d)$ is defined as

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k) \tag{2.1}$$

where $\sum_{k=1}^{K} p_k = 1$ for nonnegative mixing proportion p_k and f_k is a probability density function (pdf) of interest for the kth subpopulation parameterized by $\boldsymbol{\theta}_k$. Here $\boldsymbol{\theta}_k$ is possibly a vector of multiple parameters while $p_k \in \mathbb{R}$. For example if $f_k(\cdot)$ represents a pdf of an exponential distribution, $\boldsymbol{\theta}_k$ is a scalar: $\boldsymbol{\theta}_k \in (0, \infty) \subset \mathbb{R}$. On the other hand, when $f_k(\cdot)$ represents a pdf of a univariate normal distribution, $\boldsymbol{\theta}_k$ is a vector that consists of a mean

parameter μ and a scale parameter σ: $\boldsymbol{\theta}_k = (\mu, \sigma) \in (-\infty, \infty) \times (0, \infty) \subset \mathbb{R}^2$. In this study we assume $\boldsymbol{\theta}_k$'s are independent from each other for all $k = 1, \cdots, K$.

We assume there is a complete dataset with n observations

$$\mathcal{D}_{complete} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$$

that is independently and identically sampled from the population. Here $\mathbf{x}_i$ denotes primary data and y_i denotes a label for the subpopulation from which the primary data observed. When K subpopulations exist y_i is one of $1, \cdots, K$. In semi-supervised learning we assume the following dataset has been observed:

$$\mathcal{D}_{obs} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m), \mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\}, \quad 0 \leq m \leq n.$$

No labeled data is available if $m = 0$, while all the data are labeled if $m = n$.

It is must be noted that we assume the labeling occurs after data have been sampled from the heterogeneous population. Under the assumption we can define mechanisms of sampling each instance as follows:

1. Randomly select a mixture component k according to multinomial distribution with proportions $\{p_1, \cdots, p_K\}$.
2. Observe primary data $\mathbf{x}_i$ with probability density function $f_k(\mathbf{x}|\boldsymbol{\theta}_k)$.
3. Randomly label the instance as $y_i = k$ with some probability.

Most studies of mixtures address on maximum likelihood (ML) estimation of the parameter set

$$\Phi = (p_1, \cdots, p_K, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K),$$

as this study does. The remainder of this section introduces the existing studies for supervised, unsupervised, and semi-supervised ML estimation of FMM.

2.1.1 Supervised estimation

First we consider cases that all the observed instances are labeled, which means $\mathcal{D}_{obs}$ is exactly the same to $\mathcal{D}_{complete}$. Redner and Walker (1984) defined the log-likelihood function of

labeled observations by generalizing a two-component mixture problem (Hosmer, 1973) to K component mixture problems:

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\log\frac{n!}{\prod_{k=1}^{K}m_k!}$$

where $\mathbb{I}(\cdot)$ is an indicator function and $m_k=\sum_{i=1}^{m}\mathbb{I}(y_i=k)$. The second additive term of the above log-likelihood function is a constant term of multinomial distribution normalization which is independent to Φ. Although it is useful in comparing likelihoods of different data sets of observations, it does not provides any helpful information in comparing likelihoods for different estimates of Φ. The primary purpose of this study is to provide good estimates of Φ for given (or fixed) data set. We therefore ignore the constant term of multinomial distribution normalization in this study.

By ignoring the normalization term, the log-likelihood function of labeled samples can be derived as follows:

$$\begin{aligned}
L_s(\Phi) &= \sum_{i}^{n}\log p(\mathbf{x}_i,y_i|\Phi)\\
&= \sum_{i}^{n}\log P(y_i|\Phi)p(\mathbf{x}_i|y_i,\Phi)\\
&= \sum_{i=1}^{n}\log\sum_{k=1}^{K}\mathbb{I}(y_i=k)p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log p_k+\sum_{i=1}^{n}\sum_{k=1}^{K}I(y_i=k)\log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \qquad (2.2)
\end{aligned}$$

We can see Eq.(2.2) has an additive form of the log-likelihood of $\mathbf{p}=(p_1,\cdots,p_K)$ and the log-likelihood of $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_K)$. Eq.(2.2) is called a complete-data log-likelihood function (Dempster et al., 1977).

The maximum likelihood estimates (MLE) of Φ is a solution to the following optimization

problem:

$$\begin{aligned} \max_{\Phi} \quad & L_s(\Phi) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K \end{aligned}$$

The MLE of Φ is denoted as

$$\hat{\Phi} = (\hat{p}_1, \cdots, \hat{p}_K, \hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_K).$$

MLE $\hat{\Phi}$ can be easily and efficiently obtained by decomposing the above optimization problem into independent optimization problems. MLE of $\mathbf{p} = (p_1, \cdots, p_K)$ is a solution to

$$\begin{aligned} \max_{\mathbf{p}} \quad & \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(y_i = k) \log p_k \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K. \end{aligned}$$

As a solution to the above optimization problem, MLE of each p_k is

$$\hat{p}_k = \frac{\sum_{i=1}^{n} \mathbb{I}(y_i = k)}{n}$$

that is a proportion of instances labeled k. MLE of $\boldsymbol{\theta}_k$ is a solution to

$$\max_{\boldsymbol{\theta}_k} \quad \sum_{i=1}^{n} \mathbb{I}(y_i = k) \log f_k(\mathbf{x}_i | \boldsymbol{\theta}_k).$$

2.1.2 Unsupervised estimation

Unsupervised estimation is applied when no instance has been labeled. When the value of y_i is not given for any $1 \leq i \leq n$, the log-likelihood function has been defined (Redner and Walker, 1984) as

$$\begin{aligned} L_u(\Phi) &= \sum_{i=1}^{n} p(\mathbf{x}_i | \Phi) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k). \end{aligned} \tag{2.3}$$

The MLE of Φ is a solution to

$$\begin{aligned} \max_{\Phi} \quad & L_u(\Phi) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K. \end{aligned}$$

The logarithm of the summation of products of mixture proportions and probability densities in Eq.(2.3) makes finding MLE of Φ difficult. In contrast to the supervised estimation described in the previous section, the complicated logarithm term implies that the optimization problem cannot be decomposed into independent optimization problems.

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is one of the most widely used method to compute MLE of Φ for unsupervised estimation. Rather than directly solving the above optimization problem, the EM algorithm iteratively updates estimates of Φ with monotonically increasing likelihood values. Hasselblad (1966, 1969); Wolfe (1970); Day (1969) proposed such iterative procedures to obtain MLEs of finite mixture models. In each iteration the EM algorithm take the 'easy-to-estimate' advantages of the complete-data log-likelihood function Eq.(2.2) by solving conditional optimization problems that replaced $\mathbb{I}(y_i = k)$ with its expectation. The EM algorithms for finite mixture for unsupervised cases has been thoroughly studied, with a lot variants of types of component distributions, number of components, and constraints on mixing proportions or component distributions (Redner and Walker, 1984; McLachlan and Krishnan, 1997; McLachlan and Peel, 2000).

Let

$$\Phi^{(1)} = \left(p_1^{(1)}, \cdots, p_K^{(1)}, \boldsymbol{\theta}_1^{(1)}, \cdots, \boldsymbol{\theta}_K^{(1)}\right)$$

be an initial guess or an initial estimate of parameter Φ. Here $\boldsymbol{\theta}_1^{(1)}, \cdots, \boldsymbol{\theta}_K^{(1)}$ are supposed to be different from each other so that components of mixtures are distinguishable. The EM algorithm iteratively updates the estimates $\Phi^{(q)}$ such that the log-likelihood value monotonically increases.

At the beginning of the qth iteration ($q = 1, 2, \cdots$) of the EM algorithm, we compute a

1: Randomly initialize $\Phi^{(1)}$

2: $q \leftarrow 0$

3: **repeat**

4: $\quad q \leftarrow q + 1$

5: $\quad$ E-step: $w_{ik}^{(q)} \leftarrow \mathbb{E}\left[\mathbb{I}(y_i = k) \left| x_i, \Phi^{(q)} \right.\right]$

6: $\quad$ M-step: $\Phi^{(q+1)} \leftarrow \arg\max_{\Phi} Q(\Phi|\Phi^{(q)})$

7: **until** $L(\Phi^{(q+1)}) - L(\Phi^{(q)}) < \epsilon$

Figure 2.1: The EM algorithm for finite mixture models

conditional expectation of $\mathbb{I}(y_i = k)$ as follows:

$$\begin{aligned} w_{ik}^{(q)} &= \mathbb{E}\left[\mathbb{I}(y_i = k) \left| \mathbf{x}_i, \Phi^{(q)} \right.\right] \\ &= \frac{p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})}. \end{aligned}$$

Once we have $w_{ik}^{(q)}$ for all i and k, we define a conditional log-likelihood function

$$\begin{aligned} Q(\Phi|\Phi^{(q)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}(y_i = k)|x_i, \Phi^{(q)}\right] \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \end{aligned} \tag{2.4}$$

Let $\Phi^{(q+1)}$ be a solution to

$$\begin{aligned} \max_{\Phi} \quad & Q(\Phi|\Phi^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K. \end{aligned}$$

In EM algorithm, $L(\Phi^{(q+1)})$ is always greater than or equal to $L(\Phi^{(q)})$ (Redner and Walker, 1984). The estimates of Φ therefore keeps improving as additional iterations are performed. When the improvement is sufficiently small (i.e. $L(\Phi^{(q+1)}) - L(\Phi^{(q)}) < \epsilon$ for some $\epsilon > 0$), the iteration is terminated, and $\Phi^{(q)}$ is considered to be the MLE estimate of Φ.

Specifically, in M-step, p_k is estimated by

$$p_k^{(q+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(q)}}{n}. \tag{2.5}$$

The MLE of p_k therefore represents the expected proportion of samples that have been drawn from the kth subpopulation among given n samples. In addition $\boldsymbol{\theta}_k$ is updated by

$$\boldsymbol{\theta}_k^{(q+1)} = \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k), \tag{2.6}$$

as far as $\boldsymbol{\theta}_k$'s are independent from each other. Estimates $\boldsymbol{\theta}_k^{(q+1)}$ depend on the component distribution f_k as well as observed data $\mathbf{x}_i$. If f_k represents an exponential distribution with failure-rate parameter $\boldsymbol{\theta}_k = \lambda_k$ and $\mathbf{x}_i$ denotes right-censored lifetime (t_i, c_i) that will be described in Section 2.3.1, we have

$$\lambda_k^{(q+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(q)} c_i}{\sum_{i=1}^{n} w_{ik}^{(q)} t_i}$$

at the end of M-step in the qth iteration.

Non-identifiability is one of the critical issues in estimating FMM from only unlabeled data (McLachlan and Peel, 2000). Non-identifiability leads to multiple MLEs for a same model. For example relabeling or reordering mixture components does not affect likelihood values in unsupervised learning. Such issue also possibly happens in semi-supervised and partially supervised estimation. We do not address the non-identifiability issue in this study.

2.1.3 Semi-supervised estimation

When both labeled samples and unlabeled samples are available, either supervised estimation or unsupervised estimation is appropriate to be applied because both of them loose helpful information. Let us recall observed data

$$\mathcal{D}_{obs} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m), \mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\}, 0 \leq m \leq n.$$

With a supervised estimation approach, we only use m observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)$ while ignoring $n-m$ observations $\mathbf{x}_{m+1}, \cdots, \mathbf{x}_n$. With an unsupervised estimation approach, we use all the n observations of $\mathbf{x}$, but we ignore all the m observed labels $y_1, \cdots, y_m$. Semi-supervised estimation overcomes such limitation of uses of data.

Knowing a mechanism of observing labels is crucial in estimating FMM. To define mechanisms of label observations, we introduce an indicator variable

$$\delta_i = \begin{cases} 0 & \text{if } y_i \text{ has been reported,} \\ 1 & \text{if } y_i \text{ has been missed.} \end{cases}$$

In $\mathcal{D}_{obs}$, $\delta_1 = \cdots = \delta_m = 0$ and $\delta_{m+1} = \cdots = \delta_n = 1$. For a random sample $(\mathbf{x}, y)$ from population, let

$$\gamma_k = P(\delta = 0 | y = k)$$

be a probability of observing a label when the sample is drawn from the kth subpopulation. A vector

$$\Gamma = (\gamma_1, \cdots, \gamma_K)$$

represents a mechanism of observing labels for samples from K subpopulations.

We evaluate likelihoods of observed data based not only on Φ but also on Γ. For labeled samples $(i = 1, \cdots, m)$, a likelihood of each sample observation is defined as

$$\begin{aligned} f(\mathbf{x}_i, y_i, \delta_i = 0 | \Phi, \Gamma) &= \sum_{k=1}^{K} \mathbb{I}(y_i = k) f(\mathbf{x}_i, y_i = k, \delta_i = 0 | \Phi, \Gamma) \\ &= \sum_{k=1}^{K} \mathbb{I}(y_i = k) P(\delta_i = 0 | \Gamma) f(\mathbf{x}_i, y_i = k | \Phi) \\ &= \sum_{k=1}^{K} \mathbb{I}(y_i = k) P(\delta_i = 0 | \Gamma) P(y_i = k | \Phi) f(\mathbf{x}_i | y_i = k, \Phi) \\ &= \sum_{k=1}^{K} \mathbb{I}(y_i = k) \gamma_k p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k). \end{aligned}$$

For unlabeled samples $(i = m+1, \cdots, n)$, a likelihood of observing each sample is formulated by

$$\begin{aligned} f(\mathbf{x}_i, \delta_i = 1 | \Phi, \Gamma) &= \sum_{k=1}^{K} f(\mathbf{x}_i, y_i = k, \delta_i = 1 | \Phi, \Gamma) \\ &= \sum_{k=1}^{K} P(\delta_i = 1 | \Gamma) f(\mathbf{x}_i, y_i = k | \Phi) \\ &= \sum_{k=1}^{K} P(\delta_i = 1 | \Gamma) P(y_i = k | \Phi) f(\mathbf{x}_i | y_i = k, \Phi) \\ &= \sum_{k=1}^{K} (1 - \gamma_k) p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k). \end{aligned}$$

Therefore the log-likelihood function for semi-supervised estimation can be defined by

$$\begin{aligned} L_{ss}(\Phi,\Gamma) &= \sum_{i=1}^{m}\log f(\mathbf{x}_i,y_i,\delta_i=0|\Phi,\Gamma)+\sum_{i=m+1}^{n}\log f(\mathbf{x}_i,\delta_i=1|\Phi,\Gamma) \\ &= \sum_{i=1}^{m}\log\sum_{k=1}^{K}\mathbb{I}(y_i=k)\gamma_k p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\sum_{i=m+1}^{n}\log\sum_{k=1}^{K}(1-\gamma_k)p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \quad (2.7) \end{aligned}$$

In this section, we investigate two mechanisms of observing labels: common missing label mechanism (CML) and class-conditional missing mechanism (CCML) (Miller and Browning, 2003).

2.1.3.1 Semi-supervised estimation with common missing label (CML) assumptions

We define a mechanism of label observation as a *common missing label mechanism* if

$$\gamma_1=\cdots=\gamma_K.$$

It represents that the probability of observing labels does not depend on the origin subpopulation.

We let $\gamma=\gamma_1=\cdots=\gamma_K$. Then log-likelihood function Eq.(2.7) can be redefined by

$$\begin{aligned} L_{ss.cml}(\Phi,\gamma) &= \sum_{i=1}^{m}\log\sum_{k=1}^{K}\mathbb{I}(y_i=k)\gamma p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\sum_{i=m+1}^{n}\log\sum_{k=1}^{K}(1-\gamma)p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log\gamma p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\sum_{i=m+1}^{n}\log(1-\gamma)\sum_{k=1}^{K}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log\gamma+\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &\quad +\sum_{i=m+1}^{n}\log(1-\gamma)+\sum_{i=m+1}^{n}\log\sum_{k=1}^{K}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{m}\log\gamma+\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}(y_i=k)\log p_k+\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{I}(y_i=k)f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &\quad +\sum_{i=m+1}^{n}\log(1-\gamma)+\sum_{i=m+1}^{n}\log\sum_{k=1}^{K}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \quad (2.8) \end{aligned}$$

The MLE of γ is a proportion of labeled samples:

$$\hat{\gamma}=\frac{m}{n}.$$

To estimate MLE of Φ, the EM algorithm is also used (Miller and Browning, 2003). We start with an initial estimate $\Phi^{(1)}$. In E-step, a conditional expectation of $\mathbb{I}(y_i = k)$ at the qth iteration is computed by

$$w_{ik}^{(q)} = \begin{cases} \mathbb{I}(y_i = k) & i = 1, \cdots, m \\ \dfrac{p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})} & i = m+1, \cdots, n. \end{cases}$$

Because estimation of γ is independent from the EM algorithm, a conditional log-likelihood function to be maximized in M-step is defined as

$$Q(\Phi|\Phi^{(q)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$$

which is identical to the conditional log-likelihood function for unsupervised estimation with given $w_{ik}^{(q)}$.

2.1.3.2 Semi-supervised estimation with class-conditional missing label (CCML) assumptions

In contrast to CML mechanism, a *class-conditional missing label* mechanism represents the condition

$$\gamma_k \neq \gamma_l, \quad \exists k \neq l.$$

Therefore, we use Eq.(2.7) as a log-likelihood function. The log-likelihood function can be reformulated by

$$\begin{aligned} L_{ss,ccml}(\Phi, \Gamma) &= \sum_{i=1}^{m} \log \sum_{k=1}^{K} \mathbb{I}(y_i = k)\gamma_k p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=m+1}^{n} \log \sum_{k=1}^{K} (1-\gamma_k) p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbb{I}(y_i = k) \log \gamma_k p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=m+1}^{n} \log \sum_{k=1}^{K} (1-\gamma_k) p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbb{I}(y_i = k) \log \gamma_k + \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbb{I}(y_i = k) \log p_k \\ &\quad + \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbb{I}(y_i = k) \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=m+1}^{n} \log \sum_{k=1}^{K} (1-\gamma_k) p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \quad (2.9) \end{aligned}$$

With CCML mechanism, Γ as well as Φ should be estimated by the EM algorithm. A definition of conditional expectation $w_{ik}^{(q)}$ is

$$w_{ik}^{(q)} = \begin{cases} \mathbb{I}(y_i = k) & i = 1, \cdots, m \\ \dfrac{(1-\gamma_k)p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K}(1-\gamma_l)p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})} & i = m+1, \cdots, n. \end{cases}$$

A conditional log-likelihood function is defined as

$$\begin{aligned} Q(\Phi, \Gamma|\Phi^{(q)}, \Gamma^{(q)}) &= \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik}^{(q)} \log\gamma_k + \sum_{i=m+1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log(1-\gamma_k) \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \end{aligned}$$

$\Phi^{(q+1)}$ and $\Gamma^{(q+1)}$ be solutions to the following optimization problem:

$$\begin{aligned} \max_{\Phi,\Gamma} \quad & Q(\Phi, \Gamma|\Phi^{(q)}, \Gamma^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K \\ & 0 \leq \gamma_k \leq 1, \quad k = 1, \cdots, K. \end{aligned}$$

In contrast to CML cases, parameter Γ is updated in each M-step of the EM algorithm by

$$\gamma_k^{(q+1)} = \frac{\sum_{i=1}^{m}\mathbb{I}(y_i = k)}{\sum_{i=1}^{n} w_{ik}^{(q)}}.$$

With given $w_{ik}^{(q)}$, computation of $\Phi^{(q+1)}$ is identical to unsupervised estimation as the same to CML cases.

2.2 Partially Supervised Learning

2.2.1 Partial label

In section 2.1 we assumed observed data is either labeled or unlabeled. With binary classes ($K = 2$) they are the only possible cases we can consider. With $K > 3$ classes however there is possible observations that are neither labeled nor unlabeled. Such observed data has been called data with *partial labels* (Ambroise and Govaert, 2000; Cour et al., 2011), *ambiguous labels*

(Cour et al., 2009), or *imprecise labels* (Vannoorenberghe and Smets, 2005; Côme et al., 2008, 2009). In this study, we mostly use the terminology 'partial label' to represent such labels.

Hereafter, we redefine the complete data $\mathcal{D}_{complete}$ as

$$\mathcal{D}_{complete} = \{(\mathbf{x}_1, \mathbf{z}_1), \cdots, (\mathbf{x}_n, \mathbf{z}_n)\}$$

where $\mathbf{z}_i = (z_{i1}, \cdots, z_{iK})$ is a classification vector so that $z_{ik} = 1$ if the ith observation is sampled from the kth subpopulation and $z_{ik} = 0$ otherwise:

$$z_{ik} = \begin{cases} 1 & \text{if } i \in \text{class } k, \\ 0 & \text{otherwise.} \end{cases}$$

Because each observation is assumed to belong to exactly one subpopulation we have

$$\sum_k z_{ik} = 1, \quad \forall i. \tag{2.10}$$

Unfortunately we cannot observe all z_{ik}'s in general. For cases where values of some z_{ik}'s are unobservable partial labels are used for representing a set of candidate subpopulations to which an instance possibly belongs (Cour et al., 2009, 2011). For the ith sample, we may not observe the exact $\mathbf{z}_i$. Instead z_{ik} may be observed for some $k \in \{1, \cdots, K\}$. We define an indicator vector $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{iK})$ such that

$$\delta_{ik} = \begin{cases} 0 & \text{if } z_{ik} \text{ has been reported,} \\ 1 & \text{otherwise.} \end{cases}$$

When $\delta_{ik} = 1$ for some k, we obtain a partial label for the ith sample, which means precise information about the origin of the ith sample is missing. Such process of losing specified information about the origin is called *masking* (Usher and Hodgson, 1988; Usher and Guess, 1989; Guess et al., 1991; Lin and Guess, 1994; Ramon et al., 1995; Park, 2005; Flehinger et al., 1996, 1998). It is a special case of *coarsening* (Heitjan and Rubin, 1991; Gill et al., 1997), that is applied to nominal data. The observed partial label therefore can be considered as a coarsened version of the true label.

Due to Eq.(2.10), once we observe $\delta_{ik} = 0$ and $z_{ik} = 1$ for a certain k, we immediately know $z_{ij} = 0$ for all $j \in \{1, \cdots, K\} \backslash k$. Also if we observe $\delta_{ij} = 0$ and $z_{ij} = 0$ for all $j \in \{1, \cdots, K\} \backslash k$,

we immediately know $z_{ik} = 1$. Hence,

$$\delta_{ik} = 0 \text{ and } z_{ik} = 1 \quad \Longleftrightarrow \quad \delta_{ij} = 0 \text{ and } z_{ij} = 0 \text{ for all } j \in \{1, \cdots, K\} \backslash k.$$

Let $\tilde{\mathbf{z}}_i = (\tilde{z}_{i1}, \cdots, \tilde{z}_{iK})$ be an observed label of the ith sample so that

$$\tilde{z}_{ik} = \begin{cases} 0 & \text{if it is known that } i \notin \text{class } k, \\ 1 & \text{otherwise.} \end{cases} \tag{2.11}$$

Eq. (2.11) implies that

$$z_{ik} \leq \tilde{z}_{ik}, \quad \forall i, k. \tag{2.12}$$

A deterministic coarsening (or masking) process obtain $\tilde{\mathbf{z}}_i$ from $\mathbf{z}_i$ and $\boldsymbol{\delta}_i$ as follows:

$$\tilde{z}_{ik} = \begin{cases} 1 & \text{if } \delta_{ik} = 1, \\ z_{ik} & \text{if } \delta_{ik} = 0. \end{cases} \tag{2.13}$$

It implies that $\tilde{\mathbf{z}}_i$ completely explains the value of $\boldsymbol{\delta}_i$ as follows:

$$\boldsymbol{\delta}_i = \begin{cases} (0, \cdots, 0) & \text{if } \sum_{k=1}^{K} \tilde{z}_{ik} = 1, \\ \tilde{\mathbf{z}}_i & \text{if } \sum_{k=1}^{K} \tilde{z}_{ik} > 1. \end{cases}$$

Hereafter, a set of observed data is redefined by using partial labels as follows:

$$\mathcal{D}_{obs} = \{(\mathbf{x}_1, \tilde{\mathbf{z}}_1), \cdots, (\mathbf{x}_n, \tilde{\mathbf{z}}_n)\}.$$

2.2.1.1 Partial labeling vs. multi-labeling / soft labeling

In this study we deal with data where each instance belongs to exactly one subpopulation. Partial labels are defined as a set of candidate subpopulations that an instance possibly belongs to (Cour et al., 2009, 2011). A partial label excludes a subpopulation that is not the candidate for an instance with certainty. Therefore, the partial labels have imprecise but certain information about the membership to subpopulations.

Problems with multi-labels assume an instance can belong to multiple origins (Tsoumakas and Katakis, 2007), while subpopulations are supposed to be disjoint to each other in problems with partial labels. Multi-labels are supposed to have precise information about a set of all

the subpopulations that an instance is involved into. Therefore, the multi-labels are clearly different from the partial labels.

Partial labels are also different from uncertain labels. Uncertain labels have information about beliefs for an instance of belonging to each subpopulation (Vannoorenberghe and Denœux, 2002). Vannoorenberghe and Smets (2005), Côme et al. (2008), Côme et al. (2009), and Szczurek et al. (2010) mostly focused on the estimation of finite mixtures with the uncertain labels, which adapted fuzzy logics. The uncertain labels are, therefore, clearly different from partial labels that have information with certainty. Côme et al. (2009) stated learning from uncertain labels can generalize learning from partial labels for finite mixture models.

2.2.2 Learning FMM from partial labels

Ambroise and Govaert (2000) proposed the EM algorithm of estimating FMM from partially labeled samples. They developed the EM algorithm to maximize the following log-likelihood function:

$$L_{ps}(\Phi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \tilde{z}_{ik} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{2.14}$$

With an initial estimate $\Phi^{(1)}$, in E-step, a conditional expectation of z_{ik} at the qth iteration is computed by

$$w_{ik}^{(q)} = \frac{\tilde{z}_{ik} p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} \tilde{z}_{il} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})}.$$

A conditional log-likelihood function to be maximized at M-step is defined by

$$Q(\Phi|\Phi^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}).$$

$\Phi^{(q+1)}$ is a solution to

$$\begin{aligned} \max_{\Phi} \quad & Q(\Phi|\Phi^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K. \end{aligned}$$

In particular, p_k is estimated by

$$p_k^{(q+1)} = \frac{\sum_{i=1}^{k} w_{ik}^{(q)}}{n}.$$

With given $w_{ik}^{(q)}$ therefore the M-step is identical to the unsupervised estimation.

When handling partial labels we should carefully consider the missing mechanisms of labeling information (Lin and Guess, 1994). In Eq. (2.14) however the missing label mechanism has been ignored. In Chapter 3, we will show Ambroise and Govaert (2000)'s method implicitly assumes a specific missing label mechanism, and hence produces biased estimates given various other missing label mechanisms.

2.3 Survival Time Models

Survival time analysis is statistical analysis of time to failure of a system. Definition of failure depends on a system to be analyzed. In clinical data analysis, particularly, failure may be defined as death of a patient having a specific disease. It is very common in area of biostatistics to analyze time to death for patients. Over the past few decades numerous studies have been made on univariate and multivariate survival analysis based on continuous, nominal, and ordinal variables. In this study we only consider univariate parametric distribution to estimate distributions of time to death.

2.3.1 Univariate parametric survival time models

Let T be a continuous variable of time to death. The followings are three popular categories of functions to represent survival time patterns:

- $f(t)$: a probability density function (pdf) of time to death,
- $S(t)$: a survival function that represents $P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u$,
- $h(t)$: a hazard function that represent a local survival rate, $\frac{f(t)}{S(t)}$.

The aboves describe univariate survival time models because values of the functions only depend on time to death. If a value of T is observed for all patients in a database, $f(t)$, $S(t)$, and $h(t)$ can be estimated based on them.

In many cases exact time to death cannot be observed because of failure in tracking patients or death with causes other than of interest. Time to death is also unknown when patients are

still alive when collecting data. In such cases we use the last observed time when patients were alive, which called *right-censoring* time. To distinguish between failure time (exact time to death) and right-censoring time, the following two values must be obtained for each patient i:

- t_i: the last observed time for patient i before death,
- c_i: 1 if t_i is failure time, 0 if t_i is right-censoring time.

For cases $c_i = 0$, the only thing we know is that patient i had been survived longer than t_i.

By defining $f(t|\boldsymbol{\theta})$ as a parametric form of $f(t)$, estimating $\boldsymbol{\theta}$ is identical to estimating $f(t)$. Let the ith observation of survival time be

$$\mathbf{x}_i = (t_i, c_i).$$

for $i = 1, \cdots, n$ where n be the total number of patients of interest in a database. Then, the log-likelihood function of $\boldsymbol{\theta}$ is formulated as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(t_i|\boldsymbol{\theta})^{c_i} S(t_i|\boldsymbol{\theta})^{1-c_i}.$$

Maximum likelihood estimator of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$$

where Θ represents a solution space of $\boldsymbol{\theta}$.

An exponential distribution is one of the most popular probability distributions of failure time of a single system where no explanatory variable for the failure time is available. An exponential distribution is defined by a single parameter of a constant failure rate $\lambda \in (0, \infty)$. A pdf of the exponential distribution is defined as

$$f(t|\lambda) = \lambda e^{-\lambda t},$$

and the survival distribution is defined as

$$S(t|\lambda) = e^{-\lambda t}$$

for $t \in (0, \infty)$. The log-likelihood function is therefore defined as

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{n} \log \left(\lambda e^{-\lambda t_i}\right)^{c_i} \left(e^{-\lambda t_i}\right)^{1-c_i} \\ &= \sum_{i=1}^{n} \log \lambda^{c_i} e^{-\lambda t_i} \\ &= (\log \lambda) \sum_{i=1}^{n} c_i - \lambda \sum_{i=1}^{n} t_i. \end{aligned}$$

MLE of λ then is estimated by

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} t_i}.$$

2.3.2 FMM on univariate survival time models

FMM is widely applicable to medical data analysis (Schlattmann, 2008). In particular, if a population consists of two or more homogeneous subpopulations each survival function of which is distinguishable from the others, a distribution of failure time in a population is considered a mixture of survival functions of the subpopulations (Mendenhall and Hader, 1958; Larson and Dinse, 1985; Peng and Dear, 2000).

Mixtures of exponential distributions (MoE) have been studied to model failure time of population when it consists of two or more subpopulations, each of which is supposed to have a distinguishable failure rate from others (Mendenhall and Hader, 1958; Rider, 1961; Jewell, 1982; Papadapoulos and Padgett, 1986). A system of interest is possibly a heterogeneous system having two or more types or causes of failure. In such cases, MoE tells how much proportions the subpopulations take among the population of interest and how the time to failure is distributed within each subsystem. MoE generally gives more details of the system failure than a single exponential distribution when the system of interest is not homogeneous.

The EM algorithm (Dempster et al., 1977) guarantees convergence to the unique ML estimates of MoE parameters for fixed numbers of subsystems, when no labeling information is given to any observation (Jewell, 1982). In E-step, a conditional expectation for the ith patient of belonging to the kth class is computed by

$$w_{ik}^{(q)} = \frac{p_k \left(\lambda_k^{(q)}\right)^{c_i} e^{-\lambda_k^{(q)} t_i}}{\sum_{l=1}^{K} p_l \left(\lambda_l^{(q)}\right)^{c_i} e^{-\lambda_l^{(q)} t_i}}.$$

In M-step of the qth iteration of the EM algorithm, λ_k is estimated by

$$\lambda_k^{(q)} = \frac{\sum_{i=1}^n w_{ik}^{(q)} c_i}{\sum_{i=1}^n w_{ik}^{(q)} t_i}.$$

As described in Section 2.1–2.2, we can generalize the above EM algorithm to semi-supervised or partially supervised estimations by applying proper expectation $w_{ik}^{(q)}$.

In the area of reliability engineering, mixtures of failure time distributions with masked data have been studied since the 1980's (Miyakawa, 1984; Usher and Hodgson, 1988; Usher and Guess, 1989; Guess et al., 1991; Lin and Guess, 1994; Ramon et al., 1995; Park, 2005; Flehinger et al., 1996, 1998). In failure time analysis of machines, information that a failure of a specific component caused the system failure at time t implies that the other components have not been failed until t or have survived longer than t in other words. In the listed studies above different likelihood functions from the above ones have been designed to incorporate such additional information in estimating FMM.

This study does not assume such additional information is obtained. The reliability engineering literature assumes any of K risk factors are possibly realized in every machine. So the component led to the system failure can be observable only after the system failure actually occurred. In this study on the other hand each patient is supposed to have only one risk factor. It implies that the risk factor is observable before the death of a patient.

2.4 Summary

We have introduced detailed concepts and learning methods of finite mixture models, partially supervised learning, and survival time models that have been thoroughly studied over the last several decades. Throughout this study we combine those three concepts to learn survival time distribution of heterogeneous population from partially labeled data. An existing study for partially supervised learning of finite mixture models has been introduced in Section 2.2.2. In the next chapter we proposed more generalized partially supervised learning methods for finite mixture models.

CHAPTER 3. EM ALGORITHMS FOR ESTIMATING FINITE MIXTURE MODELS FROM PARTIAL LABELS

Here we propose the EM algorithm for finite mixture models with partial labels, which is the first part as well as the most generalized picture of this thesis.

3.1 Coarsening Probabilities

In this section we address mechanisms of observing labels that have discussed for semi-supervised learning (Section 2.1.3) to partially supervised learning problems. Let us have a set of 'pre-defined' J unique observable (or coarsened versions of) label values

$$\mathcal{C}_{\tilde{\mathbf{z}}} = \{\tilde{\mathbf{z}}_1^*, \tilde{\mathbf{z}}_2^*, \cdots, \tilde{\mathbf{z}}_J^*\}$$

where

$$\tilde{\mathbf{z}}_j^* = (\tilde{z}_{j1}^*, \cdots, \tilde{z}_{jK}^*).$$

So $\tilde{\mathbf{z}}_i \in \mathcal{C}_{\tilde{\mathbf{z}}}$ holds for all $i = 1, \cdots, n$. We call $\tilde{\mathbf{z}}_j^*$ the jth pattern of partial labels. It is important to keep in mind that $\mathcal{C}_{\tilde{\mathbf{z}}}$ is defined before observing the samples. We define all the labels that are possibly observed rather than taking the observed labels among n samples.

Let us recall the SEER data example introduced in Chapter 1. We label classes of the farthest lymph nodes involvement as 1 for `Not involved`, 2 for `Celiac/Hepatic`, 3 for `Other regional`, and 4 for `Distant`. Then Figure 3.1 defines seven label patterns $\tilde{\mathbf{z}}_1^*, \cdots, \tilde{\mathbf{z}}_7^*$ as shown in Table 3.1. Ordering classes $1, \cdots, K$ or patterns $1, \cdots, J$ does not cause differences in finding MLE of finite mixture models.

Let τ_{ij} be an indicator variable that

$$\tau_{ij} = \mathbb{I}(\tilde{\mathbf{z}}_i = \tilde{\mathbf{z}}_j^*) \quad , \forall i = 1, \cdots, n, \quad j = 1, \cdots, J. \tag{3.1}$$

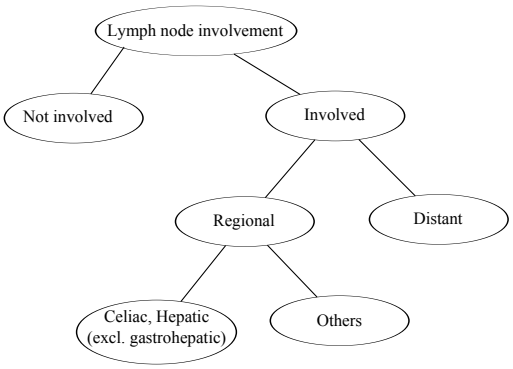


Figure 3.1: Observable values on lymph node involvement by gastric cancer tumors in SEER research data (1988–2003).

Table 3.1: Predefined 7 observable label patterns for lymph nodes involvement of gastric cancer tumor (1988–2003).

j	Description	$\tilde{z}^*_{j1}$	$\tilde{z}^*_{j2}$	$\tilde{z}^*_{j3}$	$\tilde{z}^*_{j4}$
1	Not involved	1	0	0	0
2	Celiac/Hepatic	0	1	0	0
3	Other regional	0	0	1	0
4	Distant	0	0	0	1
5	Regional	0	1	1	0
6	Involved	0	1	1	1
7	Unknown	1	1	1	1

Then the following two conditions hold:

$$\sum_{j=1}^{J} \tau_{ij} = 1 \qquad ,\forall i = 1, \cdots, n \tag{3.2}$$

$$\sum_{j=1}^{J} \tau_{ij} \tilde{z}^*_{jk} = \tilde{z}_{ik} \qquad ,\forall i, k. \tag{3.3}$$

Let γ_{jk} be a conditional probability of observing label $\tilde{\mathbf{z}}^*_j$ where the true origin is class k:

$$\begin{aligned} \gamma_{jk} &= P(\tilde{\mathbf{z}}_i = \tilde{\mathbf{z}}^*_j | z_{ik} = 1) \\ &= P(\tau_{ij} = 1 | z_{ik} = 1) \end{aligned} \tag{3.4}$$

We call γ_{jk} a coarsening probability of class k with pattern j. With γ_{jk}, vector Γ representing a mechanism of observing labels is redefined by

$$\Gamma = (\gamma_{11}, \cdots, \gamma_{JK}). \tag{3.5}$$

The following two conditions must hold for coarsening probabilities:

$$\sum_{j=1}^{J} \gamma_{jk} = 1 \quad \forall k, \tag{3.6}$$

$$\gamma_{jk} \geq 0 \quad \forall j, k. \tag{3.7}$$

In this study we assume that the observed label must be valid so that does not deliver false knowledge about the true class membership as described in Eq. (2.12):

$$z_{ik} \leq \tilde{z}_{ik} \quad \forall i, k.$$

It leads to additional condition on γ_{jk} such that

$$\gamma_{jk} = \tilde{z}^*_{jk} \gamma_{jk}, \tag{3.8}$$

which implies $\gamma_{jk} = 0$ if $\tilde{z}^*_{jk} = 0$.

3.2 Likelihood Function for Learning FMM

By using coarsening probabilities γ_{jk}, we define a sample probability density function

$$\begin{aligned}
f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi,\Gamma) &= \sum_{k=1}^{K} f(\mathbf{x}_i, \tilde{\mathbf{z}}_i, z_{ik}=1|\Phi,\Gamma) \\
&= \prod_{j=1}^{J}\left[\sum_{k=1}^{K} f(\mathbf{x}_i, \tilde{\mathbf{z}}_i=\tilde{\mathbf{z}}_j^*, z_{ik}=1|\Phi,\Gamma)\right]^{\mathbb{I}(\tilde{\mathbf{z}}_i=\tilde{\mathbf{z}}_j^*)} \\
&= \prod_{j=1}^{J}\left[\sum_{k=1}^{K} f(\mathbf{x}_i, \tau_{ij}=1, z_{ik}=1|\Phi,\Gamma)\right]^{\tau_{ij}} \\
&= \prod_{j=1}^{J}\left[\sum_{k=1}^{K} P(z_{ik}=1|\Phi) f(\mathbf{x}_i, \tau_{ij}=1|\Phi,\Gamma,z_{ik}=1)\right]^{\tau_{ij}} \\
&= \prod_{j=1}^{J}\left[\sum_{k=1}^{K} P(\tau_{ij}=1|\Gamma,z_{ik}=1)P(z_{ik}=1|\Phi) f(\mathbf{x}_i|\Phi,z_{ik}=1)\right]^{\tau_{ij}} \\
&= \prod_{j=1}^{J}\left[\sum_{k=1}^{K} \gamma_{jk} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)\right]^{\tau_{ij}}.
\end{aligned}$$

By assuming data have been independently observed or sampled the log-likelihood function on $\mathcal{D}_{obs}$ is defined by

$$L_{ps}(\Phi,\Gamma) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} \gamma_{jk} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \tag{3.9}$$

Eq. (3.9) is a log-likelihood function with consideration of coarsening probabilities which have been ignored in Ambroise and Govaert (2000); Côme et al. (2009).

As shown for semi-supervised learning cases, different assumptions on the mechanism of observing labels lead to different estimates of FMM. In this chapter, we investigate the EM algorithms to estimate FMM with the following three different mechanisms of observing labels:

- overall common missing label mechanism (OCML)
- pattern-conditional missing label mechanism (PCML)
- class-pattern-conditional missing label mechanism (CPCML)

3.3 Overall Common Missing Label mechanism (OCML)

3.3.1 FMM under OCML

OCML represents a mechanism that all the possible combinations of a coarsening pattern and a true class share a common coarsening probability. OCML implies all the coarsening probabilities γ_{jk} are the same as far as $\tilde{z}^*_{jk} = 1$:

$$\gamma_{jk} = \gamma_{hl}, \quad \forall (j,k) \neq (h,l) \text{ s.t. } \tilde{z}^*_{jk} = \tilde{z}^*_{hl} = 1. \tag{3.10}$$

Under OCML coarsening probabilities γ_{jk} can be expressed by using only a single parameter γ such that

$$\gamma_{jk} = \gamma \tilde{z}^*_{jk}, \quad \forall j,k. \tag{3.11}$$

In addition Eq. (3.6) is redefined by

$$\sum_{j=1}^{J} \gamma \tilde{z}^*_{jk} = 1, \quad \forall k. \tag{3.12}$$

Eq. (3.12) sets γ to be

$$\gamma = 1 \Big/ \sum_{j=1}^{J} \tilde{z}^*_{jk} \tag{3.13}$$

for all $k = 1, \cdots, K$. To satisfy Eq.(3.13) for all k, we must have a set of observable label patterns that

$$\sum_{j=1}^{J} \tilde{z}^*_{jk} = \sum_{j=1}^{J} \tilde{z}^*_{jl}, \quad \forall k \neq l. \tag{3.14}$$

Eq. (3.14) represents that all the classes must have the same number of possible coarsening patterns to define OCML on observing class labels. In a set of observable patterns on lymph nodes involvement attributes shown in Table 3.1, we can find that $\sum_{j=1}^{J} \tilde{z}^*_{j1} \neq \sum_{j=1}^{J} \tilde{z}^*_{j2}$. Therefore OCML is not a possible mechanism of observing lymph nodes involvement information.

3.3.2 EM-OCML: EM algorithm for learning FMM under OCML

The log-likelihood function Eq.(3.9) can be redefined under OCML by using Eq. (3.11) as follows:

$$\begin{aligned} L_{ps,ocml}(\Phi,\gamma) &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\sum_{k=1}^{K}\gamma\tilde{z}_{jk}^{*}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\log\sum_{k=1}^{K}\gamma\tilde{z}_{ik}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\log\gamma\sum_{k=1}^{K}\tilde{z}_{ik}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= n\log\gamma+\sum_{i=1}^{n}\log\sum_{k=1}^{K}\tilde{z}_{ik}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= n\log\gamma+L_{ps}(\Phi), \end{aligned} \tag{3.15}$$

where $L_{ps}(\Phi)$ has been defined in Eq. (2.14). Then MLE of FMM $\hat{\Phi}$ under OCML assumption is obtained by solving the following optimization problem:

$$\begin{aligned} \max_{\Phi,\gamma}\quad & L_{ps,ocml}(\Phi,\gamma) \\ \text{s.t.}\quad & \sum_{k=1}^{K}p_k=1 \\ & p_k\geq 0,\quad k=1,\cdots,K \\ & \sum_{j=1}^{J}\gamma\tilde{z}_{jk}^{*}=1,\quad k=1,\cdots,K. \end{aligned} \tag{3.16}$$

Because constraints on Φ and γ are independent to each other, we can independently estimate MLEs of Φ and γ. First $\hat{\gamma}$ is a solution to

$$\begin{aligned} \max_{\gamma}\quad & n\log\gamma \\ \text{s.t.}\quad & \sum_{j=1}^{J}\gamma\tilde{z}_{jk}^{*}=1,\quad k=1,\cdots,K \end{aligned}$$

The equality constraint determines MLE to be

$$\hat{\gamma}=K\Big/\sum_{j=1}^{J}\sum_{k=1}^{K}\tilde{z}_{jk}^{*} \tag{3.17}$$

if $\sum_{j=1}^{J} \tilde{z}^*_{jk} \neq \sum_{j=1}^{J} \tilde{z}^*_{jl}$ for any $k \neq l$ **then**
  return No MLE exists
end if
$\hat{\gamma} \leftarrow K / \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{z}^*_{jk}$
$q \leftarrow 0$
Initialize $\Phi^{(1)}$
repeat
  $q \leftarrow q + 1$
  E-step: $w_{ik}^{(q)} \leftarrow \tilde{z}_{ik} p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}) / \sum_{l=1}^{K} \tilde{z}_{il} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})$
  M-step(1): $p_k^{(q+1)} \leftarrow \sum_{i=1}^{k} w_{ik}^{(q)} / n$
  M-step(2): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$
until $L_{ps,ocml}(\Phi^{(q+1)}, \hat{\gamma}) - L_{ps,ocml}(\Phi^{(q)}, \hat{\gamma}) < \epsilon$
return $\hat{\Phi} \leftarrow \Phi^{(q)}$

Figure 3.2: EM-OCML: EM algorithm for learning FMM under OCML mechanism

as far as Eq. (3.14) is satisfied. In case we have a feasible $\hat{\gamma}$, we find MLE of Φ as a solution to

$$\begin{aligned} \max_{\Phi} \quad & L_{ps}(\Phi) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K. \end{aligned}$$

We have seen in Section 2.2.2 that the EM algorithm to solve the above optimization problem has been investigated by Ambroise and Govaert (2000). Although the authors did not explicitly state the mechanism of observing partial labels, they have implicitly assumed OCML in their EM algorithm. A detailed EM algorithm for learning FMM under OCML (EM-OCML) is shown in Figure 3.2.

3.4 Pattern-Conditional Missing Label Mechanism (PCML)

3.4.1 FMM under PCML

PCML has more general condition than OCML. PCML allows coarsening probabilities can be different from each other if the coarsened patterns are different. PCML however still restricts

that a chance of observing a specific pattern does not depend on the true class. It implies

$$\gamma_{jk} = \gamma_{jl}, \quad ,\forall j,\ k \neq l,\ \tilde{z}^*_{jk} = \tilde{z}^*_{jl} = 1. \tag{3.18}$$

We introduce a new parameter set

$$\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_J),$$

where

$$\gamma_j \geq 0 \quad \forall j.$$

Coarsening probabilities γ_{jk} then can be completely determined by γ_j as

$$\gamma_{jk} = \gamma_j \tilde{z}^*_{jk} \quad \forall j, k. \tag{3.19}$$

Eq.(3.6) is then redefined by replacing γ_{jk} with Eq. (3.19):

$$\sum_{j=1}^{J} \gamma_j \tilde{z}^*_{jk} = 1 \quad ,\forall k \tag{3.20}$$

It is not straightforward to define feasible regions for $\boldsymbol{\gamma}$. However we can find an example that find feasible $\boldsymbol{\gamma}$. Let us have only two possible partial labels $\tilde{\mathbf{z}}^*_1 = (1, 1, 0)$ and $\tilde{\mathbf{z}}^*_2 = (0, 1, 1)$. To satisfy Eq.(3.20) for $k \in \{1, 3\}$, we set $\gamma_1 = \gamma_2 = 1$. With this solution however $\sum_{j=1}^{2} \gamma_j z_{j2} = 2$ so that Eq.(3.20) does not holds for $k = 2$. We therefore cannot find feasible $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ with the above set of possible partial labels. On the other side, a feasible $\boldsymbol{\gamma}$ can always be found in the following two cases:

- If a set of observable partial label includes all the precise labels, we can find a feasible $\boldsymbol{\gamma}$. Let $\tilde{\mathbf{z}}^*_1 = (1, 0, \cdots, 0)$, $\tilde{\mathbf{z}}^*_2 = (0, 1, \cdots, 0), \cdots, \tilde{\mathbf{z}}^*_K = (0, 0, \cdots, K)$. Then $\gamma_1 = \cdots = \gamma_K = 1$ and $\gamma_{K+1} = \cdots = \gamma_J = 0$ represent a feasible $\boldsymbol{\gamma}$.

- If unlabeled data is observable, we can find a feasible $\boldsymbol{\gamma}$. Let $\tilde{\mathbf{z}}^*_J = (1, 1, \cdots, 1)$. Then $\gamma_1 = \cdots = \gamma_{J-1} = 0$ and $\gamma_J = 1$ is a feasible $\boldsymbol{\gamma}$.

Therefore we can find MLE of FMM under PCML for the SEER survival time data that are labeled by lymph nodes involvement (Table 3.1). The above two conditions however are not the only cases that we can find a feasible $\boldsymbol{\gamma}$. We propose an algorithm (Figure 3.3) to find a feasible $\boldsymbol{\gamma}$, if exists, to be used as an initial parameter value in the EM algorithm.

1: $K_1 \leftarrow \{k : \sum_{j=1}^{J} \tilde{z}^*_{jk} = 1\}$

2: $J_1 \leftarrow \{j : \tilde{z}^*_{jk} = 1, k \in K_1\}$

3: **while** $K_1 \neq \{1, \cdots, K\}$ **do**

4: **if** $J_1 = \{1, \cdots, J\}$ **then**

5: **return** No feasible initial $\boldsymbol{\gamma}$

6: **end if**

7: $J_2 \leftarrow \{j : \sum_{k \notin K_1} \tilde{z}^*_{jk} \geq 1\}$

8: $h \leftarrow \arg\min_{j \in J_2} \sum_{k \notin K_1} \tilde{z}^*_{jk}$

9: $J_1 \leftarrow J_1 \cup h$

10: $K_1 \leftarrow K_1 \cup \{k : \tilde{z}^*_{hk} = 1\}$

11: **end while**

12: $\gamma_j \leftarrow 1 / \max_k \sum_{l=1}^{J} \tilde{z}^*_{lk}$ for $j \notin J_1$

13: γ_j for $j \in J_1$ are determined by γ_j for $j \notin J_1$ $\{J - |J_1|$ is the number of parameters on $\boldsymbol{\gamma}\}$

Figure 3.3: An algorithm to find a feasible $\boldsymbol{\gamma}$ under PCML

3.4.2 EM-PCML: EM algorithm for learning FMM under PCML

By using new parameter set $\boldsymbol{\gamma}$ the log-likelihood function Eq.(3.9) can be redefined under PCML as

$$L_{ps,pcml}(\Phi, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} \gamma_j \tilde{z}^*_{jk} p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k).$$

MLE of FMM under PCML assumption is then the answer to

$$\begin{aligned} \max_{\Phi, \boldsymbol{\gamma}} \quad & L_{ps,pcml}(\Phi, \boldsymbol{\gamma}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad \forall k \\ & \sum_{j=1}^{J} \gamma_j \tilde{z}^*_{jk} = 1, \quad \forall k \\ & \gamma_j \geq 0 \quad , \forall j. \end{aligned} \tag{3.21}$$

The EM algorithm iteratively find MLEs of both Φ and $\boldsymbol{\gamma}$ under PCML assumption (Figure 3.4). In E-step of the qth iteration of the EM algorithm, a conditional expectation of z_{ik}

with current estimates $\Phi^{(q)}$ and $\boldsymbol{\gamma}^{(q)}$ is computed by

$$\begin{aligned} w_{ik}^{(q)} &= \mathbb{E}\left[z_{ik} \,\middle|\, \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)}\right] \\ &= P(z_{ik}=1|\mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)}) \\ &= \frac{f(z_{ik}=1, \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)})}{f(\mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)})} \\ &= \prod_{j=1}^{J}\left[\frac{\gamma_j^{(q)}\tilde{z}_{jk}^* p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K}\gamma_j^{(q)}\tilde{z}_{jl}^* p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})}\right]^{\tau_{ij}}, \end{aligned} \tag{3.22}$$

where $0^0 = 1$. We define a conditional log-likelihood function on PCML as

$$\begin{aligned} Q_{ps,pcml}(\Phi, \boldsymbol{\gamma}|\Phi^{(q)}, \boldsymbol{\gamma}^{(q)}) &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \gamma_j \tilde{z}_{jk}^* p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \gamma_j \tilde{z}_{jk}^* + \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \gamma_j \tilde{z}_{jk}^* + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \gamma_j \tilde{z}_{jk}^* + Q(\Phi|\Phi^{(q)}) \end{aligned} \tag{3.23}$$

Eq.(3.23) shows the conditional log-likelihood function is composed of a log-likelihood for Φ and a log-likelihood for $\boldsymbol{\gamma}$. In M-step, we find an estimate of Φ that maximize

$$\begin{aligned} \max_{\Phi} \quad & Q(\Phi|\Phi^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad k = 1, \cdots, K \end{aligned}$$

which we have seen in Section 2.1.2. To find an estimate of $\boldsymbol{\gamma}$, we solve the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\gamma}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \gamma_j \tilde{z}_{jk}^* \\ \text{s.t.} \quad & \sum_{j=1}^{J} \gamma_j \tilde{z}_{jk}^* = 1, \quad k = 1, \cdots, K \\ & \gamma_j \geq 0 \quad , \forall j = 1, \cdots, J \end{aligned} \tag{3.24}$$

1: $q \leftarrow 0$

2: Initialize $\boldsymbol{\gamma}^{(1)}$ by using an algorithm in Figure 3.3.

3: **if** A feasible $\boldsymbol{\gamma}^{(1)}$ cannot be found **then**

4: **return** No MLE exists

5: **end if**

6: Initialize $\Phi^{(1)}$

7: **repeat**

8: $q \leftarrow q + 1$

9: E-step: $w_{ik}^{(q)} \leftarrow \sum_{j=1}^{J} \tau_{ij} \left[\gamma_j^{(q)} \tilde{z}_{jk}^* p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}) / \sum_{l=1}^{K} \gamma_j^{(q)} \tilde{z}_{jl}^* p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)}) \right]$

10: M-step(1): $p_k^{(q+1)} \leftarrow \sum_{i=1}^{k} w_{ik}^{(q)} / n$

11: M-step(2): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$

12: M-step(3): Obtain $\boldsymbol{\gamma}^{(q+1)}$ by solving Eq. (3.24) using BFGS method.

13: **until** $L_{ps,pcml}(\Phi^{(q+1)}, \boldsymbol{\gamma}^{(q+1)}) - L_{ps,pcml}(\Phi^{(q)}, \boldsymbol{\gamma}^{(q)}) < \epsilon$

14: **return** $\hat{\Phi} \leftarrow \Phi^{(q)}$, $\hat{\boldsymbol{\gamma}} \leftarrow \boldsymbol{\gamma}^{(q)}$

Figure 3.4: EM-PCML: EM algorithm for learning FMM under PCML mechanism

Although it is hard to find a closed form of $\boldsymbol{\gamma}^{(q+1)}$, we can use nonlinear optimization methods like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to obtain $\boldsymbol{\gamma}^{(q+1)}$ that is a solution to Eq. (3.24).

3.5 Class-Pattern-Conditional Missing Label Mechanism (CPCML)

3.5.1 FMM under CPCML

CPCML represents the most general mechanism on observing partial labels in this study. CPCML implies the coarsening probabilities do not have any restrictions besides Eqs.(3.6)–(3.8). It implies coarsening probabilities can be different from each other with respect to the true class as well as the observed patterns. The parameters γ_{jk}'s are not further simplified with smaller number of parameters.

We can find a feasible estimate of FMM under CPCML as far as

$$\sum_{j=1}^{J} z_{jk}^* \geq 1, \tag{3.25}$$

for all $k = 1, \cdots, K$. If Eq. (3.25) does not hold for any k, $\sum_{j=1}^{J} \gamma_{jk} = 0$ due to Eq. (3.8), so that Eq. (3.6) must be violated.

3.5.2 EM-CPCML: EM algorithm for learning FMM under CPCML

By incorporating Eq. (3.8) into the log-likelihood function Eq.(3.9), we have a log-likelihood function under CPCML that

$$L_{ps,cpcml}(\Phi, \Gamma) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} \tilde{z}_{jk}^{*} \gamma_{jk} p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k). \tag{3.26}$$

MLEs of Φ and Γ represent a solution to the following optimization problem:

$$\begin{aligned} \max_{\Phi, \Gamma} \quad & L_{ps,cpcml}(\Phi, \Gamma) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad \forall k \\ & \sum_{j=1}^{J} \gamma_{jk} = 1, \quad \forall k \\ & \gamma_{jk} \geq 0, \quad \forall j, k. \end{aligned} \tag{3.27}$$

Let us have current estimates $\Phi^{(q)}$ and $\Gamma^{(q)}$ at starting the qth iteration of the EM algorithm. In E-step, a conditional expectation of z_{ik} is computed by

$$\begin{aligned} w_{ik}^{(q)} &= \mathbb{E}\left[z_{ik} \middle| \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)} \right] \\ &= \prod_{j=1}^{J} \left[\frac{\gamma_{jk}^{(q)} p_k^{(q)} f_k(\mathbf{x}_i | \boldsymbol{\theta}_k^{(q)})}{\sum_{l=1}^{K} \gamma_{jl}^{(q)} p_l^{(q)} f_l(\mathbf{x}_i | \boldsymbol{\theta}_l^{(q)})} \right]^{\tau_{ij}} \end{aligned} \tag{3.28}$$

which $\gamma_j^{(q)} \tilde{z}_{jk}^{*}$ in Eq.(3.22) is replaced with $\gamma_{jk}^{(q)}$. With the conditional expectation, we define a

conditional log-likelihood function

$$\begin{aligned} Q_{ps,cpcml}(\Phi,\Gamma|\Phi^{(q)},\Gamma^{(q)}) &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*}\gamma_{jk}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*}\gamma_{jk} + \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*}\gamma_{jk} + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*}\gamma_{jk} + Q(\Phi|\Phi^{(q)}) \end{aligned} \tag{3.29}$$

In M-step, we find an estimate of Φ to maximize $Q(\Phi|\Phi^{(q)})$ in the same ways we have shown for unsupervised, semi-supervised, or partially supervised learning under OCML or PCML. An estimate of Γ must be a solution to

$$\begin{aligned} \max_{\Gamma} \quad & \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*}\gamma_{jk} \\ \text{s.t.} \quad & \sum_{j=1}^{J}\gamma_{jk} = 1, \quad \forall k \\ & \gamma_{jk} \geq 0, \quad \forall j,k. \end{aligned}$$

Hence,

$$\gamma_{jk}^{(q+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}w_{ik}^{(q)}}{\sum_{i=1}^{n} w_{ik}^{(q)}}. \tag{3.30}$$

A detailed EM algorithm for learning finite mixture models under CPCML assumption is described in Figure 3.5.

3.6 Comparison of Log-Likelihood under OCML, PCML, and CPCML

Sections 3.3–3.5 have described three models for FMM learning from partially labeled data. In this section we describe how such three models are connected to each other from the perspective of optimization. To clearly show relations among three models, we reformulate optimization problems to find MLE of FMM under OCML (Eq. 3.16), PCML (Eq. 3.21), and CPCML (Eq. 3.27) by using the log-likelihood function

$$L_{ps}(\Phi,\Gamma) = \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\sum_{k=1}^{K}\gamma_{jk}p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$$

1: $q \leftarrow 0$

2: Initialize $\Gamma^{(1)}$ with $\gamma_{jk}^{(1)} = 0$ if $\tilde{z}_{jk}^{*} = 0$.

3: Initialize $\Phi^{(1)}$.

4: **repeat**

5: $\quad q \leftarrow q + 1$

6: $\quad$ E-step: $w_{ik}^{(q)} \leftarrow \sum_{j=1}^{J} \tau_{ij} \left[\gamma_{jk}^{(q)} p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}) / \sum_{l=1}^{K} \gamma_{jl}^{(q)} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)}) \right]$

7: $\quad$ M-step(1): $p_k^{(q+1)} \leftarrow \sum_{i=1}^{k} w_{ik}^{(q)} / n$

8: $\quad$ M-step(2): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$

9: $\quad$ M-step(3): $\gamma_{jk}^{(q+1)} \leftarrow \sum_{i=1}^{n} \tau_{ij} w_{ik}^{(q)} / \sum_{i=1}^{n} w_{ik}^{(q)}$

10: **until** $L_{ps,cpcml}(\Phi^{(q+1)}, \Gamma^{(q+1)}) - L_{ps,cpcml}(\Phi^{(q)}, \Gamma^{(q)}) < \epsilon$

11: **return** $\hat{\Phi} \leftarrow \Phi^{(q)}$, $\hat{\Gamma} \leftarrow \Gamma^{(q)}$

Figure 3.5: EM-CPCML: EM algorithm for learning FMM under CPCML mechanism

that has been defined in Eq. (3.9). Under all the three missing label mechanisms, the following constraints must hold:

$$\left.\begin{aligned} &\sum_{k=1}^{K} p_k = 1, \\ &p_k \geq 0, \quad \forall k, \end{aligned}\right\} \text{ common proportion constraints}$$

$$\left.\begin{aligned} &\sum_{j=1}^{J} \gamma_{jk} = 1, \quad \forall k, \\ &\gamma_{jk} \geq 0, \quad \forall j, k, \\ &\gamma_{jk} = \tilde{z}_{jk}^{*} \gamma_{jk}, \quad \forall j, k, \end{aligned}\right\} \text{ common coarsening constraints}$$

where common coarsening constraints have been described in Section 3.1.

An optimization problem to find MLE of FMM under CPCML (Eq. 3.27) is formulated as

$$\begin{aligned} \max_{\Phi,\Gamma} \quad & L_{ps}(\Phi, \Gamma) \\ \text{s.t.} \quad & \text{common proportion constraints} \\ & \text{common coarsening constraints.} \end{aligned} \tag{3.31}$$

Under PCML we have additional constraints from Eq. (3.18) that

$$\gamma_{jk} = \gamma_{jl}, \quad \forall j = 1, \cdots, J, k \neq l : \tilde{z}_{jk}^{*} = \tilde{z}_{jl}^{*} = 1.$$

So MLE of FMM under PCML is an optimal solution to

$$\begin{aligned} \max_{\Phi,\Gamma} \quad & L_{ps}(\Phi,\Gamma) \\ \text{s.t.} \quad & \text{common proportion constraints} \\ & \text{common coarsening constraints} \\ & \text{Eq. (3.18)}. \end{aligned} \tag{3.32}$$

OCML has even more strict constraints on coarsening probabilities from Eq. (3.10) such that

$$\gamma_{jk} = \gamma_{hl}, \quad \forall (j,k) \neq (h,l) : \tilde{z}^*_{jk} = \tilde{z}^*_{hl} = 1.$$

MLE of FMM under OCML is therefore an optimal solution to

$$\left.\begin{aligned} \max_{\Phi,\Gamma} \quad & L_{ps}(\Phi,\Gamma) \\ \text{s.t.} \quad & \text{common proportion constraints} \\ & \text{common coarsening constraints} \\ & \text{Eq. (3.10)}. \end{aligned}\right\} \tag{3.33}$$

Optimization problems Eqs.(3.31)–(3.33) share a common objective function $L_{ps}(\Phi,\Gamma)$. So the only differences among three models are in solution spaces. Eq.(3.32) and Eq.(3.33) have additional constraints compared to Eq.(3.31). Solution spaces of FMM under PCML and OCML are therefore included into a solution space of FMM under CPCML; it means that CPCML leads to the highest optimal log-likelihood value among three models. In addition, it is clear that Eq. (3.10) always satisfies Eq. (3.18), while such relation does not hold in the opposite way. Hence we know that a solution space of FMM under OCML is a part of a solution space of FMM under PCML. On the same objective function, an optimal solution from a larger solution space guarantees as good solution as an optimal solution from a smaller solution space that is a subset of the larger one. Therefore in partially supervised learning of finite mixture models,

$$L_{ps,cpcml}(\hat{\Phi},\hat{\Gamma}) \geq L_{ps,pcml}(\hat{\Phi},\hat{\boldsymbol{\gamma}}) \geq L_{ps,ocml}(\hat{\Phi},\hat{\gamma}), \tag{3.34}$$

where MLEs of FMM are exists for all the three missing label mechanisms.

3.7 Model Selection

The previous section has shown that a solution space of an optimization problem under CPCML includes solutions spaces under OCML and PCML; it has been caused by additional equality constraints under OCML and PCML. In statistics, it implies that OCML and PCML represent reduced models of CPCML. Section 3.4 shown that only J parameters are needed to represent PCML mechanisms. Under OCML we only need one as shown in Section 3.3, which dramatically reduces the number of parameters when J and K are large. The complexity of FMM is the highest under CPCML, the second highest under PCML, and the lowest under OCML. So MLE of FMM from EM-OCML is expected to be the most precise, while MLE from EM-CPCML is the least. In cases OCML or PCML holds in observed data therefore EM-CPCML includes unnecessarily many parameters and sacrifices precision of estimation without benefits against EM-OCML or EM-PCML. It is called an over-fitting problem. On the other side, strict equality constraints on coarsening mechanisms, Eq. (3.10) and Eq. (3.18), may sacrifice accuracy of MLEs in cases such equality constraints do not hold in observed data. Falsely defined assumptions generally lead to biased MLEs from the true models.

Knowing missing label mechanisms in observed data is therefore crucial to preserve precision of MLE as much as possible without sacrificing accuracy. In many cases that we unfortunately do not know the underlying missing label mechanisms, it may be useful to evaluate how the underlying missing label mechanism is plausible to be OCML, PCML, or CPCML after obtaining MLEs under each missing label mechanism assumption. Model selection represents a such task of choosing the most appropriate statistical model from several candidates. We suggest Akaike information criterion (AIC) (Akaike, 1974) to be used for model selection. AIC is defined as

$$AIC = -2(\text{log-likelihood}) + 2(\text{number of parameters}).$$

Here log-likelihood represents a value of log-likelihood function with MLE of FMM. Also the number of parameters used in computing AIC represents the least number of parameters to be estimated in learning FMM from observed data. In this study we have two sets of parameters: Φ and Γ. Because proportion constraints on Φ are the same for EM-OCML, EM-PCML and EM-CPCML, the number of parameters in Φ does not affect differences in AIC among models with

different missing label mechanisms. On the other side, constraints on Γ are different between models as shown in Eqs. (3.31)–(3.33), which causes differences in the number of parameters to be estimated.

The complete Γ is a set of $J \times K$ coarsening probabilities γ_{jk} (Eq. 3.5). However constraints on Γ predetermine some γ_{jk} before using the observed data. Under OCML Γ can be represented by using only one coarsening probability γ (Eq. 3.11). Moreover MLE of γ is not estimated by observed data but determined by predefined observable patterns z^*_{jk} (Eq. 3.13). Hence γ does not contribute on AIC. AIC for MLE under OCML is therefore computed by

$$AIC_{ps,ocml} = -2L_{ps,ocml}(\hat{\Phi}, \hat{\gamma}) + 2|\Phi|, \tag{3.35}$$

where $\hat{\Phi}$ and $\hat{\gamma}$ are MLEs of Φ and γ, respectively, and $|\Phi|$ represents the number of parameters to be estimated in Φ.

Under CPCML, Eq.(3.6) implies that a coarsening probability for one pattern is explained by coarsening probabilities for the other patterns within each class. In addition Eq.(3.8) implies that the value of coarsening probability γ_{jk} is predetermined to be zero if $z^*_{jk} = 0$. The number of parameters on Γ is therefore

$$\sum_{k=1}^{K} \left(\sum_{j=1}^{J} z^*_{jk} - 1 \right).$$

We therefore define AIC for MLE under CPCML as

$$AIC_{ps,cpcml} = -2L_{ps,cpcml}(\hat{\Phi}, \hat{\Gamma}) + 2|\Phi| + 2\sum_{k=1}^{K} \left(\sum_{j=1}^{J} z^*_{jk} - 1 \right). \tag{3.36}$$

Under PCML Γ can be represented by using J coarsening probabilities in $\boldsymbol{\gamma}$ (Eq. 3.19). The least number of parameters to represent $\boldsymbol{\gamma}$ has been obtained at the end of the algorithm in Figure 3.3 as $J - |J_1|$. AIC for FMM under PCML is therefore defined as

$$AIC_{ps,pcml} = -2L_{ps,pcml}(\hat{\Phi}, \hat{\Gamma}) + 2|\Phi| + 2J - 2|J_1|, \tag{3.37}$$

where J_1 is obtained by the algorithm in Figure 3.3.

Generally the missing label mechanism corresponding to the least AIC value is considered as the most plausible one. We therefore select a MLE from the missing label mechanism that shows the minimum AIC value among $AIC_{ps,ocml}$, $AIC_{ps,pcml}$ and $AIC_{ps,cpcml}$ as the best estimate on $\mathcal{D}_{obs}$.

3.8 Summary

In this chapter we have proposed EM algorithms to estimate finite mixture models by incorporating partial labels. We have considered three missing label mechanisms: overall common missing label mechanism (OCML), pattern-conditional missing label mechanism (PCML), and class-pattern-conditional missing label mechanism (CPCML). CPCML represented the most general missing label mechanism, while OCML and PCML have restricted solution spaces of FMM. For cases of unknown underlying missing label mechanism on observed data, we suggested AIC to be used for model selection criteria.

We have not restricted patterns of partial labels in this chapter. In the following chapter we particularly focus on a taxonomic system that hierarchically specifies observable patterns of partial labels.

CHAPTER 4. LEARNING FINITE MIXTURE MODELS FROM ATTRIBUTE VALUE TAXONOMY

In the previous chapter, we investigated the estimation of finite mixture models from partially labeled data without any restriction on the patterns of observable labels. In this chapter, we introduce attribute value taxonomy (AVT) as a restriction on the observed labels of subpopulations and study how to use AVT-guided data for learning finite mixture models.

4.1 Attribute Value Taxonomy

Expert knowledge about useful hierarchical categorization of population may guide data input procedures. For example SEER database hierarchically defines the levels of lymph nodes involvement for gastric cancer (Figure 4.1). Gastric cancer cases are split into cases with lymph nodes involved by tumor and cases without lymph nodes involved. Cases that lymph nodes involved by tumor are further investigated whether the tumor has been spread to lymph nodes in distant organs far away from the stomach or has been spread within nearby or regional lymph nodes. For cases with only regional lymph nodes involved, it is investigated whether celiac lymph nodes or hepatic lymph nodes are involved or not. Such specifications of the level of lymph nodes involvement are important to estimate a person's prognosis like survival time. From Figure 4.1 we know celiac and hepatic lymph nodes are regional lymph nodes for gastric cancer, not distant lymph nodes. We also know not all regional lymph nodes are either celiac or hepatic lymph nodes; there are chances that other regional lymph nodes are involved by gastric cancer tumor.

Attribute value taxonomy (AVT) is a tree of attribute values that represents such 'is-a' relationships among values on a nominal attribute (Almuallim et al., 1995; Zhang, 2005;

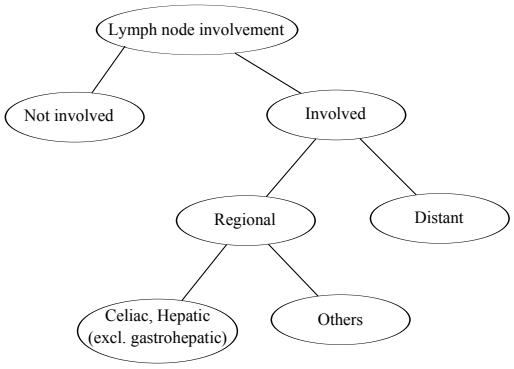


Figure 4.1: AVT applied to SEER research data for lymph node involvement of gastric cancer tumors.

Sharma and Poole, 2005). The root node of AVT represents unspecified value on the attribute. Branches in AVT represent specification of attribute values from a relatively imprecise value to disjoint further detailed values. So on a single path from the top-most or the root node to a bottom or a leaf node in AVT, the farther from the root node, the more specific information contained about the attribute. Values corresponding to leaf nodes of AVT are desired to be observed because they contain the most specific information on the attribute. However values on the attributes may not be completely specified due to lack of information or uncontrolled data input practices. AVT limits observable values on the attribute to leaf and inner nodes of AVT. With AVT in Figure 4.1 for example SEER allows people to label a gastric cancer case as `Regional` if it is certain that the case is neither `Not involved` nor `Distant`. There is still an ambiguity in an answer to which of `Celiac/Hepatic` or `Other regional` best represents the case, but it is allowed ambiguity to be stored in SEER database. On the other hand an imprecise label representing a case is either `Celiac/Hepatic` or `Distant` is not allowed to be input.

4.2 Learning Finite Mixture Models with Alternative Formulation

In this section we propose an alternative parameterization for learning finite mixture models. This section will deliver more intuitive explanation about learning finite mixture models for

whom are not familiar with coarsening probabilities in Section 3.1.

4.2.1 Parameterization of finite mixture models

Let us recall definitions of data that have been defined in Sections 2.2 and 3.1. A probability density function of values of interest $\mathbf{x}$ is supposed to be a mixture of K components,

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k).$$

The objective of this study is find MLE of $f(\mathbf{x})$ by utilizing all observed information about the component origin from which each data was observed. A data set with n complete observations is defined by

$$\mathcal{D}_{complete} = \{(\mathbf{x}_1, \mathbf{z}_1), \cdots, (\mathbf{x}_n, \mathbf{z}_n)\}$$

where $\mathbf{z}_i = (z_{i1}, \cdots, z_{iK})$ and

$$z_{ik} = \begin{cases} 1 & \text{if } i \in \text{class } k, \\ 0 & \text{otherwise.} \end{cases}$$

Element values of an indicator vector $\boldsymbol{\delta}_i = (\delta_{i1}, \cdots, \delta_{iK})$ are set to be

$$\delta_{ik} = \begin{cases} 0 & \text{if } z_{ik} \text{ has been reported,} \\ 1 & \text{otherwise.} \end{cases}$$

Then an observed label of the ith sample $\tilde{\mathbf{z}}_i = (\tilde{z}_{i1}, \cdots, \tilde{z}_{iK})$ is valued as

$$\tilde{z}_{ik} = \begin{cases} 1 & \text{if } \delta_{ik} = 1, \\ z_{ik} & \text{if } \delta_{ik} = 0. \end{cases}$$

So rather than $\mathcal{D}_{complete}$ we commonly observe a data set

$$\mathcal{D}_{obs} = \{(\mathbf{x}_1, \tilde{\mathbf{z}}_1), \cdots, (\mathbf{x}_n, \tilde{\mathbf{z}}_n)\}.$$

With predefined J unique observable label patterns $\tilde{\mathbf{z}}_1^*, \tilde{\mathbf{z}}_2^*, \cdots, \tilde{\mathbf{z}}_J^*$, we define an indicator variable τ_{ij} so that

$$\tau_{ij} = \mathbb{I}(\tilde{\mathbf{z}}_i = \tilde{\mathbf{z}}_j^*) \quad , \forall i = 1, \cdots, n, \quad j = 1, \cdots, J.$$

Modeling a sample density function

$$\begin{aligned} f(\mathbf{x}_i, \tilde{\mathbf{z}}_i) &= \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} f(\mathbf{x}_i, \tau_{ij}=1, z_{ik}=1) \\ &= \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} f(\mathbf{x}_i|\tau_{ij}=1, z_{ik}=1) P(\tau_{ij}=1, z_{ik}=1) \\ &= \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} f(\mathbf{x}_i|z_{ik}=1) P(\tau_{ij}=1, z_{ik}=1) \\ &= \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) P(\tau_{ij}=1, z_{ik}=1) \end{aligned}$$

is a crucial task to obtain MLE of $f(\mathbf{x})$. Particularly there are multiple ways to model a joint probability of a true class and an observed label $P(\tau_{ij}=1, z_{ik}=1)$. In Chapter 3 we parameterized $P(\tau_{ij}=1|z_{ik}=1)$ by γ_{jk} and $P(z_{ik}=1)$ by p_k to represent the joint probability such that

$$\begin{aligned} P(\tau_{ij}=1, z_{ik}=1) &= P(\tau_{ij}=1|z_{ik}=1) P(z_{ik}=1) \\ &= \gamma_{jk} p_k. \end{aligned}$$

In this section we parameterize $P(z_{ik}=1|\tau_{ij}=1)$ and $P(\tau_{ij}=1)$ to represent the same joint probability; it leads to more intuitive EM algorithm for learning FMM from AVT-guided data. Let we define parameters p_{jk} and π_j that

$$p_{jk} = P(z_{ik}=1|\tau_{ij}=1), \tag{4.1}$$

$$\pi_j = P(\tau_{ij}=1). \tag{4.2}$$

p_{jk} represents a proportion of class k within instances that are labeled as $\tilde{\mathbf{z}}_j^*$. In addition π_j denotes a marginal probability that instances are labeled as $\tilde{\mathbf{z}}_j^*$. Then the joint probability of a true class and an observed label is represented by

$$\begin{aligned} P(\tau_{ij}=1, z_{ik}=1) &= P(\tau_{ij}=1) P(z_{ik}=1|\tau_{ij}=1) \\ &= \pi_j p_{jk}. \end{aligned}$$

So we can learn FMM with new parameter sets

$$\begin{aligned} \mathbf{P} &= (p_{11}, \cdots, p_{JK}), \\ \Pi &= (\pi_1, \cdots, \pi_J), \end{aligned}$$

with the following constraints:

$$\begin{aligned}
\sum_{k=1}^{K} p_{jk} &= 1, \quad \forall j, && (4.3)\\
p_{jk} &\geq 0, \quad \forall j,k, && (4.4)\\
\sum_{j=1}^{J} \pi_j &= 1, && (4.5)\\
\pi_j &\geq 0, \quad \forall j. && (4.6)
\end{aligned}$$

As similar to constraints on γ_{jk}, the relation between the observed label and the true class membership

$$z_{ik} \leq \tilde{z}_{ik} \quad \forall i,k,$$

leads to additional condition on p_{jk} such that

$$p_{jk} = \tilde{z}^*_{jk} p_{jk}, \tag{4.7}$$

which implies $p_{jk} = 0$ if $\tilde{z}^*_{jk} = 0$.

We still need to have mixture parameters Φ for representing mixture components $\boldsymbol{\theta}_k$, but mixture proportions p_k can be completely explained by the new parameters because

$$\begin{aligned}
p_k &= P(z_{ik} = 1) \\
&= \sum_{j=1}^{J} P(\tau_{ij} = 1) P(z_{ik} = 1 | \tau_{ij} = 1) \\
&= \sum_{j=1}^{J} \pi_j p_{jk}. && (4.8)
\end{aligned}$$

So a marginal pdf of $\mathbf{x}$ is defined as

$$\begin{aligned}
f(\mathbf{x}) &= \sum_{k=1}^{K} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k) \\
&= \sum_{k=1}^{K} \sum_{j=1}^{J} p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k) \\
&= \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K} p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k) \\
&= \sum_{j=1}^{J} \pi_j f_{(j)}(\mathbf{x}), && (4.9)
\end{aligned}$$

where

$$f_{(j)}(\mathbf{x}) = \sum_{k=1}^{K} p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k). \tag{4.10}$$

Eq.(4.10) represents that a value of $\mathbf{x}$ with label $\tilde{\mathbf{z}}_j^*$ is supposed to be drawn from a mixture of K probability density function with different mixing proportions from a mixture on other labels as well as the overall mixture $f(\mathbf{x})$, while component parameter $\boldsymbol{\theta}_k$ is independent to the observed label.

By using new parameter sets, we define a sample probability density function

$$\begin{aligned}
f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi, \mathbf{P}, \Pi) &= \sum_{k=1}^{K} f(\mathbf{x}_i, \tilde{\mathbf{z}}_i, z_{ik} = 1|\Phi, \mathbf{P}, \Pi) \\
&= \prod_{i=1}^{J} \left[\sum_{k=1}^{K} f(\mathbf{x}_i, \tau_{ij} = 1, z_{ik} = 1|\Phi, \mathbf{P}, \Pi) \right]^{\tau_{ij}} \\
&= \prod_{i=1}^{J} \left[\sum_{k=1}^{K} P(\tau_{ij} = 1, z_{ik} = 1|\mathbf{P}, \Pi) f(\mathbf{x}_i|\Phi, z_{ik} = 1) \right]^{\tau_{ij}} \\
&= \prod_{i=1}^{J} \left[\sum_{k=1}^{K} P(\tau_{ij} = 1|\Pi) P(z_{ik} = 1|\mathbf{P}, \tau_{ij} = 1) f(\mathbf{x}_i|\Phi, z_{ik} = 1) \right]^{\tau_{ij}} \\
&= \prod_{i=1}^{J} \left[\sum_{k=1}^{K} \pi_j p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \right]^{\tau_{ij}}.
\end{aligned} \tag{4.11}$$

So a log-likelihood function for FMM is defined as follows:

$$\begin{aligned}
L_{ps}(\Phi, \mathbf{P}, \Pi) &= \sum_{i=1}^{n} \log f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi, \mathbf{P}, \Pi) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} \pi_j p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \pi_j.
\end{aligned} \tag{4.12}$$

MLE of FMM is a solution to $\arg\max_{\Phi, \mathbf{P}, \Pi} L_{ps}(\Phi, \mathbf{P}, \Pi)$ with constraints Eqs. (4.3)–(4.7) as well as Eq. (4.8).

4.2.2 Alternative EM algorithm for learning FMM under CPCML

By incorporating Eq. (4.7) into log-likelihood function Eq. (4.12), we have a new log-likelihood function under CPCML such that

$$L_{ps,cpcml}(\Phi, \mathbf{P}, \Pi) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} z_{jk}^{*} p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \pi_j. \tag{4.13}$$

Then MLE of FMM $\hat{\Phi}$ under CPCML assumption is obtained by solving the following optimization problem:

$$\begin{aligned} \max_{\Phi,\mathbf{P},\Pi} \quad & L_{ps,cpcml}(\Phi, \mathbf{P}, \Pi) \\ \text{s.t.} \quad & \sum_{j=1}^{J} \pi_k = 1 \\ & \pi_j \geq 0, \quad \forall j \\ & \sum_{k=1}^{K} p_{jk} = 1, \quad \forall j \\ & p_{jk} \geq 0, \quad \forall j, k, \\ & p_k = \sum_{j=1}^{J} \pi_j p_{jk}, \quad \forall k. \end{aligned} \tag{4.14}$$

The last constraint is nonlinear, which makes difficult to get a closed form solution to Eq. (4.14). Hence we propose an EM algorithm to obtain MLE of FMM with the new parameterization.

At the qth iteration in EM algorithm, let us have current estimates $\Phi^{(q)}$, $\mathbf{P}^{(q)}$, and $\Pi^{(q)}$. In E-step, a conditional expectation of z_{ik} is computed by

$$\begin{aligned} w_{ik}^{(q)} &= \mathbb{E}\left[z_{ik} \middle| \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \mathbf{P}^{(q)}, \Pi^{(q)} \right] \\ &= \mathbb{E}\left[z_{ik} \middle| \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \mathbf{P}^{(q)} \right] \\ &= \sum_{j=1}^{J} \tau_{ij} P(z_{ik} = 1 | \mathbf{x}_i, \tau_{ij} = 1, \Phi^{(q)}, \mathbf{P}^{(q)}) \\ &= \sum_{j=1}^{J} \tau_{ij} \frac{f(\mathbf{x}_i, z_{ik} = 1 | \tau_{ij} = 1, \Phi^{(q)}, \mathbf{P}^{(q)})}{f(\mathbf{x}_i | \tau_{ij} = 1, \Phi^{(q)}, \mathbf{P}^{(q)})} \\ &= \sum_{j=1}^{J} \tau_{ij} \frac{p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{l=1}^{K} p_{jl} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l)} \\ &= \prod_{j=1}^{J} \left[\frac{p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{l=1}^{K} p_{jl} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l)} \right]^{\tau_{ij}}. \end{aligned} \tag{4.15}$$

From Eq. (4.15) we see that marginal pattern probability Π does not affect a conditional expectation of z_{ik} to be computed within EM algorithm. It implies that Π can be estimated independently to Φ and $\mathbf{P}$, which makes the estimation simpler. MLE of Π is estimated by counting the observed label patterns:

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}, \quad \forall j = 1, \cdots, J. \tag{4.16}$$

Hence the log-likelihood function we need to maximize by using EM algorithm under CPCML is

$$L_{ps,cpcml}(\Phi, \mathbf{P}, \hat{\Pi}) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} z^*_{jk} p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) + \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \hat{\pi}_j,$$

where the second additive term is fixed by Eq. (4.16). By ignoring the second additive term of Eq. (4.17), we define a conditional log-likelihood function

$$Q_{ps,cpcml}(\Phi, \mathbf{P}|\Phi^{(q)}, \mathbf{P}^{(q)}) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} w_{ik}^{(q)} \log z^*_{jk} p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \tag{4.17}$$

where $w_{ik}^{(q)}$ is defined by Eq. (4.15). In M-step of EM algorithm, we find new estimates $\Phi^{(q+1)}$ and $\mathbf{P}^{(q+1)}$ of Φ and $\mathbf{P}$, respectively, as optimal solutions to

$$\begin{aligned} \max_{\Phi,\mathbf{P}} \quad & Q_{ps,cpcml}(\Phi, \mathbf{P}|\Phi^{(q)}, \mathbf{P}^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_{jk} = 1, \quad \forall j \\ & p_{jk} \geq 0, \quad \forall j, k, \\ & p_k = \sum_{j=1}^{J} \hat{\pi}_j p_{jk}. \end{aligned} \tag{4.18}$$

In maximizing conditional log-likelihood $Q_{ps,cpcml}(\Phi, \mathbf{P}|\Phi^{(q)}, \mathbf{P}^{(q)})$, we do not have nonlinear constraints because a parameter π_j in Eq. (4.14) has been replaced with a fixed estimate $\hat{\pi}_j$. So we can easily have closed forms of updated parameters that

$$p_{jk}^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ij} w_{ik}^{(q)}}{\sum_{i=1}^{n} \tau_{ij}}, \quad \forall j, k, \tag{4.19}$$

$$p_k^{(q+1)} = \sum_{j=1}^{J} \hat{\pi}_j p_{jk}^{(q+1)}, \quad \forall k. \tag{4.20}$$

1: $\hat{\pi}_j \leftarrow \sum_{i=1}^{n} \tau_{ij}/n$

2: $q \leftarrow 0$

3: Initialize $\mathbf{P}^{(1)}$ with $p_{jk}^{(1)} = 0$ if $\tilde{z}_{jk}^{*} = 0$.

4: Initialize $\Phi^{(1)}$ with $p_{k}^{(1)} = \sum_{j=1}^{J} \hat{\pi}_j p_{jk}^{(1)}$.

5: **repeat**

6: $q \leftarrow q + 1$

7: E-step: $w_{ik}^{(q)} \leftarrow \sum_{j=1}^{J} \tau_{ij} \left[p_{jk}^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}) / \sum_{l=1}^{K} p_{jl}^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)}) \right]$

8: M-step(1): $p_{jk}^{(q+1)} \leftarrow \sum_{i=1}^{n} \tau_{ij} w_{ik}^{(q)} / \sum_{i=1}^{n} \tau_{ij}$

9: M-step(2): $p_{k}^{(q+1)} \leftarrow \sum_{j=1}^{J} \hat{\pi}_j p_{jk}^{(q+1)}$

10: M-step(3): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$

11: **until** $L_{ps,cpcml}(\Phi^{(q+1)}, \mathbf{P}^{(q+1)}, \hat{\Pi}) - L_{ps,cpcml}(\Phi^{(q)}, \mathbf{P}^{(q)}, \hat{\Pi}) < \epsilon$

12: **return** $\hat{\Phi} \leftarrow \Phi^{(q)}$, $\hat{\mathbf{P}} \leftarrow \mathbf{P}^{(q)}$

Figure 4.2: Alternative EM-CPCML

We terminate the iteration when $[L_{ps,cpcml}(\Phi^{(q+1)}, \mathbf{P}^{(q+1)}, \hat{\Pi}) - L_{ps,cpcml}(\Phi^{(q)}, \mathbf{P}^{(q)}, \hat{\Pi})]$ is less than a predetermined threshold ϵ. A detailed EM algorithm for learning FMM under CPCML assumption with the new parameterization is described in Figure 4.2.

4.2.3 Alternative EM algorithm for learning FMM under PCML

PCML needs additional equality constraints on p_{jk} such that

$$p_{jk} = \frac{\tilde{z}_{jk}^{*} p_k}{\sum_{l=1}^{K} \tilde{z}_{jl}^{*} p_l}, \quad \forall j, k. \tag{4.21}$$

Eq. (4.21) represents that p_{jk} are determined by the overall K component mixture proportions $p_1, \cdots, p_K$. So $\mathbf{P}$ is a redundant parameter set if Φ exists. By removing $\mathbf{P}$ from parameter sets, we have a sample probability density function that

$$f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi, \Pi) = \prod_{i=1}^{J} \left[\sum_{k=1}^{K} \pi_j \frac{\tilde{z}_{jk}^{*} p_k}{\sum_{l=1}^{K} \tilde{z}_{jl}^{*} p_l} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \right]^{\tau_{ij}} \tag{4.22}$$

and a log-likelihood function that

$$\begin{aligned} L_{ps,pcml}(\Phi,\Pi) &= \sum_{i=1}^{n}\log f(\mathbf{x}_i,\tilde{\mathbf{z}}_i|\Phi,\Pi) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\sum_{k=1}^{K}\pi_j\frac{\tilde{z}^*_{jk}p_k}{\sum_{l=1}^{K}\tilde{z}^*_{jl}p_l}f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\sum_{k=1}^{K}\frac{\tilde{z}^*_{jk}p_k}{\sum_{l=1}^{K}\tilde{z}^*_{jl}p_l}f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\pi_j. \end{aligned} \tag{4.23}$$

The following relations between p_k and π_j must hold under PCML:

$$\sum_{j=1}^{J}\tilde{z}^*_{jk}\pi_j \geq p_k, \quad \forall k=1,\cdots,K, \tag{4.24}$$

$$\sum_{k=1}^{K}\tilde{z}^*_{jk}p_k \geq \pi_j, \quad \forall j=1,\cdots,J. \tag{4.25}$$

Eq. (4.24) describes that the probability of observing a label including class k should be larger than the probability of observing an instance from class k at each observation. In addition Eq. (4.25) implies that the probability of observing a specific label pattern should be less than the probability of observing an instance from one of classes that the label pattern covers. Eqs. (4.24)–(4.24) make π_j cannot be estimated independently from the other parameters as it does under CPCML. So, Π must be estimated within EM algorithm as well as Φ.

At the qth iteration in EM algorithm, let us have current estimates $\Phi^{(q)}$ and $\Pi^{(q)}$. In E-step, a conditional expectation of z_{ik} is computed by

$$\begin{aligned} w^{(q)}_{ik} &= \mathbb{E}\left[z_{ik}\left|\mathbf{x}_i,\tilde{\mathbf{z}}_i,\Phi^{(q)},\Pi^{(q)}\right.\right] \\ &= \prod_{j=1}^{J}\left[\frac{\frac{\tilde{z}^*_{jk}p_k}{\sum_{h=1}^{K}\tilde{z}^*_{jh}p_h}f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{l=1}^{K}\frac{\tilde{z}^*_{jl}p_l}{\sum_{h=1}^{K}\tilde{z}^*_{jh}p_h}f_l(\mathbf{x}_i|\boldsymbol{\theta}_l)}\right]^{\tau_{ij}} \\ &= \prod_{j=1}^{J}\left[\frac{\tilde{z}^*_{jk}p_kf_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{l=1}^{K}\tilde{z}^*_{jl}p_lf_l(\mathbf{x}_i|\boldsymbol{\theta}_l)}\right]^{\tau_{ij}}. \end{aligned} \tag{4.26}$$

In M-step we need to update estimates of Φ and Π that maximize a conditional log-likelihood function

$$Q_{ps,pcml}(\Phi,\Pi|\Phi^{(q)},\Pi^{(q)}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\sum_{k=1}^{K}w^{(q)}_{ik}\log\frac{\tilde{z}^*_{jk}p_k}{\sum_{l=1}^{K}\tilde{z}^*_{jl}p_l}f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)+\sum_{i=1}^{n}\sum_{j=1}^{J}\tau_{ij}\log\pi_j. \tag{4.27}$$

Maximizing Eq. (4.27) should be conducted with constraints Eqs. (4.24)–(4.25) as well as

$$\begin{aligned} \sum_{k=1}^{K} p_k &= 1 \\ p_k &\geq 0, \quad \forall k. \end{aligned}$$

The maximization of the conditional log-likelihood function does not look simpler or easier than that in Section 4.2.3. Therefore, we follow the EM algorithm in Section to estimate FMM under PCML.

4.3 EM Algorithm for AVT-guided Data

Now we specify the proposed EM algorithms for AVT-guided data. In this section we directly use the AVT tree structure rather than explicitly generating binary pattern variables $\tilde{z}^*_{jk}$.

4.3.1 Notation

Let us start with the example in Figure 4.1. By indexing classes as

- 1: Not involved,
- 2: Celiac/Hepatic,
- 3: Other regional,
- 4: Distant,

we define seven observable patterns $\tilde{\mathbf{z}}^*_1, \cdots, \tilde{\mathbf{z}}^*_7$ as shown in Table 4.1. By indexing nodes with $j = 1, \cdots, 7$, we develop a simplified tree $\mathcal{T}$ Figure 4.3.

General operators of trees are defined for AVT as follows:

- $Nodes(\mathcal{T})$: a set of nodes of $\mathcal{T}$,
- $Leaf(\mathcal{T})$: a set of leaf nodes of $\mathcal{T}$,
- $Root(\mathcal{T})$: the root node of $\mathcal{T}$.

Table 4.1: Observable label patterns $\tilde{\mathbf{z}}_j^*$ for lymph nodes involvement of gastric cancer tumor corresponding to Figure 4.1.

j	Description	$\tilde{z}_{j1}^*$	$\tilde{z}_{j2}^*$	$\tilde{z}_{j3}^*$	$\tilde{z}_{j4}^*$
1	Not involved	1	0	0	0
2	Celiac/Hepatic	0	1	0	0
3	Other regional	0	0	1	0
4	Distant	0	0	0	1
5	Regional	0	1	1	0
6	Involved	0	1	1	1
7	Unknown	1	1	1	1

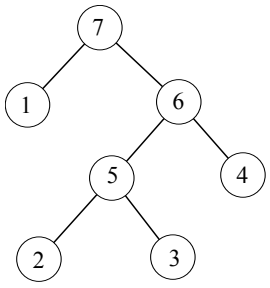


Figure 4.3: A simplified AVT $\mathcal{T}$ with node indexes.

In addition, for $j \in Nodes(\mathcal{T})$, the following operators are defined:

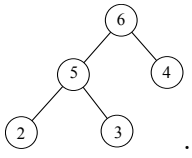
- $Anc(j, \mathcal{T})$: a set of ancestors of j,
- $Desc(j, \mathcal{T})$: a set of descendants of j,
- $Child(j, \mathcal{T})$: a set of child nodes of j,
- $Subtree(j, \mathcal{T})$: a subtree of $\mathcal{T}$ which consists of $Desc(j, \mathcal{T})$ with a root node j.

Additionally we define a depth of AVT be the maximum level of specification such that

$$Depth(\mathcal{T}) = \max_{k \in Leaf(\mathcal{T})} |Anc(k, \mathcal{T})| + 1.$$

With $\mathcal{T}$ that represented in Figure 4.3 for example

$$\begin{aligned} Nodes(\mathcal{T}) &= \{1, \cdots, 7\}, \\ Leaf(\mathcal{T}) &= \{1, 2, 3, 4\}, \\ Root(\mathcal{T}) &= 7, \\ Anc(6, \mathcal{T}) &= \{7\}, \\ Desc(6, \mathcal{T}) &= \{2, 3, 4, 5\}, \\ Child(6, \mathcal{T}) &= \{4, 5\}, \end{aligned}$$

$Subtree(6, \mathcal{T}) =$ 

.

For convenience, hereafter we index the K leaf nodes to be $1, \cdots, K$ and the root node to be J without loss of generality. Then the followings are satisfied:

$$\begin{aligned} Nodes(T) &= \{1, \cdots, J\}, \\ Leaf(T) &= \{1, \cdots, K\}, \\ Root(T) &= J. \end{aligned}$$

In addition for any node j the following relations between tree operators are generally hold:

$$\begin{aligned} Root(Subtree(j,\mathcal{T})) &= j, \\ Nodes(Subtree(j,\mathcal{T})) &= Desc(j,\mathcal{T}) \cup \{j\}, \\ Desc(j,\mathcal{T}) &= \bigcup_{l \in Child(j,\mathcal{T})} Nodes(Subtree(l,\mathcal{T})). \end{aligned}$$

Also sibling nodes l and h from the same parent j satisfy the following conditions that make subtrees disjoint from each other:

$$\begin{aligned} Nodes(Subtree(l,\mathcal{T})) \cap Nodes(Subtree(h,\mathcal{T})) &= \emptyset \\ Leaf(Subtree(l,\mathcal{T})) \cap Leaf(Subtree(h,\mathcal{T})) &= \emptyset. \end{aligned}$$

To shorten notations in the following sections we let

$$\begin{aligned} Nodes(j,\mathcal{T}) &= Nodes(Subtree(j,\mathcal{T})), \\ Leaf(j,\mathcal{T}) &= Leaf(Subtree(j,\mathcal{T})). \end{aligned}$$

4.3.2 Learning finite mixture models with AVT under CPCML

With defined AVT $\mathcal{T}$, a pdf of $\mathbf{x}$ corresponding to node j where the observed label pattern is $\tilde{\mathbf{z}}_j^*$ is defined by

$$f_{(j)}(\mathbf{x}) = \sum_{k \in Leaf(j,\mathcal{T})} p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k). \tag{4.28}$$

In addition constraint Eq. (4.3) is specified by

$$\sum_{k \in Leaf(j,\mathcal{T})} p_{jk} = 1, \quad \forall j. \tag{4.29}$$

It implies that values of $\mathbf{x}$ within each label pattern are distributed by finite mixture models. Therefore learning FMM from AVT-guided data is learning FMM within each node on AVT while all the nodes share the same component distribution parameters $\boldsymbol{\theta}_k$.

A marginal pdf of $\mathbf{x}$ is defined as

$$f_{\mathcal{T}}(\mathbf{x}) = \sum_{j \in Nodes(\mathcal{T})} \pi_j f_{(j)}(\mathbf{x}). \tag{4.30}$$

Our objective is finding MLE of $f_{\mathcal{T}}(\mathbf{x})$ by utilizing all the observed data. So the log-likelihood function to be maximized is newly defined by

$$\begin{aligned} L_{\mathcal{T},cpcml}(\Phi, \mathbf{P}, \Pi) &= \sum_{i=1}^{n} \sum_{j \in Nodes(\mathcal{T})} \tau_{ij} \log \pi_j f_{(j)}(\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \sum_{j \in Nodes(\mathcal{T})} \tau_{ij} \pi_j \log \sum_{k \in Leaf(j,\mathcal{T})} p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k). \end{aligned} \tag{4.31}$$

As same to Section 4.2, MLE of Π is obtained by setting the partial derivative of Eq. (4.31) on π_j to be zero, so that

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \tau_{ij}, \tag{4.32}$$

since $\sum_{j \in Nodes(\mathcal{T})} \tau_{ij} = n$. Therefore we define a conditional log-likelihood to be maximized within M-step of the EM algorithm as

$$Q_{\mathcal{T},cpcml}(\Phi, \mathbf{P}|\Phi^{(q)}, \mathbf{P}^{(q)}) = \sum_{i=1}^{n} \sum_{j \in Nodes(\mathcal{T})} \tau_{ij} \sum_{k \in Leaf(j,\mathcal{T})} w_{ik}^{(q)} \log p_{jk} f_k(\mathbf{x}|\boldsymbol{\theta}_k) \tag{4.33}$$

where

$$w_{ik}^{(q)} = \frac{\sum_{j \in Anc(k,\mathcal{T}) \cup \{k\}} \tau_{ij} p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{j \in Anc(k,\mathcal{T}) \cup \{k\}} \tau_{ij} \sum_{l \in Leaf(j,\mathcal{T})} p_{jl} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l)}. \tag{4.34}$$

Updating estimates of p_{jk} and $\boldsymbol{\theta}_k$ is identical to Section 4.2 with newly defined $w_{ik}^{(q)}$. The relation between p_k and p_{jk} is defined as

$$p_k = \hat{\pi}_k + \sum_{j \in Anc(k,\mathcal{T})} \hat{\pi}_j p_{jk}, \quad \forall k. \tag{4.35}$$

The EM algorithm is described in Figure 4.4.

4.3.3 Learning finite mixture models with AVT under PCML

Under PCML mechanism, relations between marginal proportions of subpopulation p_k and pattern-conditional proportions of subpopulation p_{jk} have been defined in Eq.(4.21). The relations are specified on AVT as

$$p_{jk} = \mathbb{I}(k \in Leaf(j, \mathcal{T})) \frac{p_k}{\sum_{l \in Leaf(j,\mathcal{T})} p_l}, \quad \forall j, k. \tag{4.36}$$

Eq. (4.36) represents that p_{jk} are determined by the overall K component mixture proportions $p_1, \cdots, p_K$. So $\mathbf{P}$ is a redundant parameter set if Φ exists. By Eq. (4.36) a pattern-conditional

1: $\hat{\pi}_j \leftarrow \sum_{i=1}^{n} \tau_{ij}/n$

2: $q \leftarrow 0$

3: Initialize $\mathbf{P}^{(1)}$ with $p_{jk}^{(1)} = 0$ if $k \notin \mathit{Leaf}(j, \mathcal{T})$.

4: Initialize $\Phi^{(1)}$ with $p_k^{(1)} = \sum_{j=1}^{J} \hat{\pi}_j p_{jk}^{(1)}$.

5: **repeat**

6: $\quad q \leftarrow q + 1$

7: $\quad$ E-step: $w_{ik}^{(q)} \leftarrow \sum_{j \in Anc(k,\mathcal{T}) \cup \{k\}} \tau_{ij} \left[p_{jk} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) / \sum_{l \in Leaf(j,\mathcal{T})} p_{jl} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l) \right]$

8: $\quad$ M-step(1): $p_{jk}^{(q+1)} \leftarrow \sum_{i=1}^{n} \tau_{ij} w_{ik}^{(q)} / \sum_{i=1}^{n} \tau_{ij}$

9: $\quad$ M-step(2): $p_k^{(q+1)} \leftarrow \hat{\pi}_k + \sum_{j \in Anc(k,\mathcal{T})} \hat{\pi}_j p_{jk}$

10: $\quad$ M-step(3): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$

11: **until** $L_{\mathcal{T},cpcml}(\Phi^{(q+1)}, \mathbf{P}^{(q+1)}, \hat{\Pi}) - L_{\mathcal{T},cpcml}(\Phi^{(q)}, \mathbf{P}^{(q)}, \hat{\Pi}) < \epsilon$

12: **return** $\hat{\Phi} \leftarrow \Phi^{(q)}$, $\hat{\mathbf{P}} \leftarrow \mathbf{P}^{(q)}$

Figure 4.4: Alternative EM-CPCML on AVT

pdf of $\mathbf{x}$ in Eq. (4.10) is redefined as

$$f_{(j)}(\mathbf{x}) = \frac{\sum_{k \in Leaf(j,\mathcal{T})} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k)}{\sum_{k \in Leaf(j,\mathcal{T})} p_k}. \tag{4.37}$$

For $j \notin \mathit{Leaf}(\mathcal{T})$ in particular Eq. (4.37) is specified as

$$\begin{aligned} f_{(j)}(\mathbf{x}) &= \frac{\sum_{l \in Child(j,\mathcal{T})} \sum_{k \in Leaf(l,\mathcal{T})} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k)}{\sum_{k \in Leaf(j,\mathcal{T})} p_k} \\ &= \sum_{l \in Child(j,\mathcal{T})} \left[\frac{\sum_{k \in Leaf(l,\mathcal{T})} p_k}{\sum_{k \in Leaf(j,\mathcal{T})} p_k} \right] \left[\frac{\sum_{k \in Leaf(l,\mathcal{T})} p_k f_k(\mathbf{x}|\boldsymbol{\theta}_k)}{\sum_{k \in Leaf(l,\mathcal{T})} p_k} \right] \\ &= \sum_{l \in Child(j,\mathcal{T})} \left[\frac{\sum_{k \in Leaf(l,\mathcal{T})} p_k}{\sum_{k \in Leaf(j,\mathcal{T})} p_k} \right] f_{(l)}(\mathbf{x}). \end{aligned} \tag{4.38}$$

Therefore FMM corresponding to an inner node j is a mixture of FMMs corresponding to child nodes under PCML, where the mixing proportions are determined by AVT structure as well as overall subpopulation proportion p_k.

The following relations between p_k and π_j must hold under PCML on AVT:

$$\pi_k + \sum_{j \in Anc(k,\mathcal{T})} \pi_j \geq p_k, \quad \forall k = 1, \cdots, K, \tag{4.39}$$

$$\sum_{k \in Leaf(j,\mathcal{T})} p_k \geq \pi_j, \quad \forall j = 1, \cdots, J.. \tag{4.40}$$

4.3.4 Further discussion about PCML

A pdf of $\mathbf{x}$ within a subtree of AVT can be defined by

$$\begin{aligned} f_{Subtree(j,\mathcal{T})}(\mathbf{x}) &= \frac{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l f_{(l)}(\mathbf{x})}{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l} \\ &= \sum_{k\in Leaf(j,\mathcal{T})}\frac{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk}}{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l} f_k(\mathbf{x}|\boldsymbol{\theta}_k), \end{aligned} \tag{4.41}$$

which describes that a pdf of $\mathbf{x}$ within a subtree is a mixture of pdfs corresponding to leaf nodes in the subtree with proportion $\frac{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk}}{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l}$. Under PCML, we know

$$\frac{p_{lk}}{\sum_{h\in Leaf(j,\mathcal{T})} p_{lh}} = \frac{p_k}{\sum_{h\in Leaf(j,\mathcal{T})} p_h} \tag{4.42}$$

for $l \in Anc(j,\mathcal{T})$ and $k \in Leaf(j,\mathcal{T})$. Then for $k \in Leaf(j,\mathcal{T})$

$$\begin{aligned} p_k &= \sum_{l\in Nodes(\mathcal{T})}\pi_l p_{lk} \\ &= \sum_{l\in Anc(j,\mathcal{T})}\pi_l p_{lk} + \sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk} \\ &= \sum_{l\in Anc(j,\mathcal{T})}\pi_l \frac{p_k}{\sum_{h\in Leaf(l,\mathcal{T})} p_h} + \sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk}, \end{aligned} \tag{4.43}$$

so that

$$\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk} = p_k\left(1 - \sum_{l\in Anc(j,\mathcal{T})}\frac{\pi_l}{\sum_{h\in Leaf(l,\mathcal{T})} p_h}\right). \tag{4.44}$$

Therefore a proportion of component $k \in Leaf(j,\mathcal{T})$ within FMM $f_{Subtree(j,\mathcal{T})}(\mathbf{x})$ is

$$\begin{aligned} \frac{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk}}{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l} &= \frac{\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lk}}{\sum_{h\in Leaf(j,\mathcal{T})}\sum_{l\in Nodes(j,\mathcal{T})}\pi_l p_{lh}} \\ &= \frac{p_k}{\sum_{h\in Leaf(j,\mathcal{T})} p_h}. \end{aligned} \tag{4.45}$$

From Eqs. (4.37), (4.41) and (4.45) we can know that

$$f_{(j)}(\mathbf{x}) = f_{Subtree(j,\mathcal{T})}(\mathbf{x}) \tag{4.46}$$

under PCML on AVT.

By replacing $f_{(l)}(\mathbf{x})$ with $f_{Subtree(l,\mathcal{T})}(\mathbf{x})$ in Eq.(4.38), we have

$$f_{(j)}(\mathbf{x}) = \sum_{l\in Child(j,\mathcal{T})}\left[\frac{\sum_{k\in Leaf(l,\mathcal{T})} p_k}{\sum_{k\in Leaf(j,\mathcal{T})} p_k}\right] f_{Subtree(l,\mathcal{T})}(\mathbf{x}) \tag{4.47}$$

for $j \notin \mathit{Leaf}(\mathcal{T})$. So PCML implies that FMM corresponding to an inner node j is a mixture of FMMs corresponding to subtrees whose root nodes are child nodes of the inner node. In this mixture, contributions of subtrees on the mixture are proportional to the size of subpopulation p_k corresponding to the subtrees. This proportional constraint is relaxed in the new missing label mechanism that we propose in the following section.

4.4 Learning FMM under Hierarchy-Conditional Missing Label Mechanism (HCML)

In this section we propose another missing label mechanism that we name hierarchy-conditional missing label (HCML) mechanism; it is a missing label mechanism only definable on AVT. HCML represents a limited mechanism compared to CPCML but a generalized mechanism compared to PCML. We therefore introduce HCML for robust estimation of FMMs when PCML is violated but CPCML is too general to specify missing label mechanisms on given data.

4.4.1 Hierarchy-conditional missing label mechanism (HCML)

Let us imagine that specifications of labeling information are conducted step-by-step from the root node to the leaf nodes on AVT. HCML implies that a single step specification of labeling information is conducted depending on its results but not on the following specification step. A pattern-conditional pdf of $\mathbf{x}$ under HCML is defined by

$$f_{(j)}(\mathbf{x}) = \begin{cases} f_k(\mathbf{x}|\boldsymbol{\theta}_k) & \text{if } j \in \mathit{Leaf}(\mathcal{T}), \\ \displaystyle\sum_{l \in \mathit{Child}(j,\mathcal{T})} \xi_{jl} f_{\mathit{Subtree}(l,\mathcal{T})}(\mathbf{x}) & \text{if } j \notin \mathit{Leaf}(\mathcal{T}), \end{cases} \tag{4.48}$$

where ξ_{jl} denotes a proportion of subtree $\mathit{Subtree}(l, \mathcal{T})$ within node j, $f_{\mathit{Subtree}(l,\mathcal{T})}(\mathbf{x})$ is defined in Eq. (4.41), and $\mathit{Leaf}(\mathcal{T}) = \{1, \cdots, K\}$. Under PCML we set ξ_{jl} to be $\frac{\sum_{k \in \mathit{Leaf}(l,\mathcal{T})} p_k}{\sum_{k \in \mathit{Leaf}(j,\mathcal{T})} p_k}$ as described in Eq. (4.47). Under HCML we parameterize ξ_{jl} to allow flexibility of FMM estimates in cases the underlying missing label mechanism on observed data violates PCML assumption. Now we have a new set of parameters

$$\Xi = (\xi_{K+1,1}, \cdots, \xi_{K+1,J-1}, \cdots, \xi_{J1}, \cdots, \xi_{J,J-1}).$$

The reason we do not define ξ_{jl} for $j = 1, \cdots, K$ is that nodes $1, \cdots, K$ are leaf nodes so that they do not have any child nodes in our notations. In addition ξ_{jl} is not defined for $l = J$ because node J denotes the root node so that no nodes can have node J as a child node. Ξ must satisfy the following three constraints:

$$\sum_{l \in Child(j,\mathcal{T})} \xi_{jl} = 1, \quad j = K+1, \cdots, J, \tag{4.49}$$

$$\xi_{jl} \geq 0, \quad j = K+1, \cdots, J, \ l = 1, \cdots, J-1, \tag{4.50}$$

$$\xi_{jl} = \mathbb{I}(l \in Child(j,\mathcal{T})) \cdot \xi_{jl}, \quad j = K+1, \cdots, J, \ l = 1, \cdots, J-1. \tag{4.51}$$

So it is obvious that

$$\xi_{jl} = 0, \quad \text{if } l \notin Child(j,\mathcal{T}).$$

Then the following condition must hold for $j \notin Leaf(\mathcal{T})$ and $k = 1, \cdots, K$ under HCML:

$$\begin{aligned} p_{jk} &= P(z_{ik} = 1 | \tau_{ij} = 1) \\ &= \sum_{l \in Child(j,\mathcal{T})} \xi_{jl} P\left(z_{ik} = 1 \,\middle|\, \sum_{h \in Nodes(l,h)} \tau_{ih} = 1 \right) \\ &= \sum_{l \in Child(j,\mathcal{T})} \xi_{jl} \frac{\sum_{h \in Nodes(l,\mathcal{T})} \pi_h p_{hk}}{\sum_{h \in Nodes(l,\mathcal{T})} \pi_h}. \end{aligned}$$

Therefore p_{jk} is completely determined by ξ_{jl} for $l \in Child(j,\mathcal{T})$ as well as p_{hk} for $h \in Desc(j,\mathcal{T})$. Our labeling strategy on AVT sets p_{kk} to be 1 for $k = 1, \cdots, K$ because a leaf node k represents completely specified label on class k. It implies that all p_{jk} are completely determined by multiplying all ξ_{jl}'s on a path from an inner node j to a leaf node k

$$p_{jk} = \prod_{o=1}^{O-1} \xi_{l_o l_{o+1}}, \tag{4.52}$$

where $l_o = j$, $l_O = k$, $l_{o+1} \in Child(l_o,\mathcal{T})$, and $l_o \in Anc(k,\mathcal{T})$ for $o = 1, \cdots, O-1$. So we can remove $\mathbf{P}$ from our parameter sets by taking Ξ as a parameter set. In addition p_k is also completely determined Ξ and Π by replacing p_{jk} with Eq. (4.52) in computing

$$p_k = \sum_{j=1}^{J} \pi_j p_{jk}.$$

By using new parameter set Ξ, we define a sample probability density function

$$\begin{aligned} f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi, \Pi, \Xi) &= \sum_{j=1}^{J} \tau_{ij} f(\mathbf{x}_i, \tau_{ij} = 1|\Phi, \Pi, \Xi) \\ &= \sum_{j=1}^{J} \tau_{ij} \pi_j f_{(j)}(\mathbf{x}_i|\Phi, \Xi) \\ &= \prod_{j=1}^{J} \left[\pi_j f_{(j)}(\mathbf{x}_i|\Phi, \Xi)\right]^{\tau_{ij}}, \end{aligned} \tag{4.53}$$

where $f_{(j)}(\mathbf{x}|\Phi, \Xi)$ denotes a pattern-conditional pdf $f_{(j)}(\mathbf{x})$ that is defined by Eq. (4.48). So a log-likelihood function for FMM is defined by incorporating a constraint Eq. (4.51) as follows:

$$\begin{aligned} L_{\mathcal{T},hcml}(\Phi, \Pi, \Xi) &= \sum_{i=1}^{n} \log f(\mathbf{x}_i, \tilde{\mathbf{z}}_i|\Phi, \Pi, \Xi) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \pi_j f_{(j)}(\mathbf{x}_i|\Phi, \Xi) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} \tau_{ij} \log f_j(\mathbf{x}_i|\boldsymbol{\theta}_j) + \\ &\quad \sum_{i=1}^{n} \sum_{j=K+1}^{J} \tau_{ij} \log \pi_j \sum_{l \in Child(j,\mathcal{T})} \xi_{jl} f_{Subtree(l,\mathcal{T})}(\mathbf{x}), \end{aligned} \tag{4.54}$$

where $f_{Subtree(l,\mathcal{T})}(\mathbf{x})$ is defined in Eq. (4.41).

4.4.2 EM-HCML: EM algorithm for learning FMM under HCML

The recursive multiplication of ξ_{jl} is however not intuitive for studying estimation. We therefore recall a parameter set $\Gamma = (\gamma_{jk})$ that has been defined in Chapter 3 as

$$\gamma_{jk} = P(\tau_{ij} = 1|z_{ik} = 1).$$

HCML defines an equality constraint that

$$\gamma_{jk} = \gamma_{jl}, \quad \text{if } \exists h \in Child(j, \mathcal{T}) \text{ such that } k, l \in Leaf(h, \mathcal{T}). \tag{4.55}$$

Therefore MLE of FMM under HCML is defined as an optimal solution to

$$\begin{aligned}
\max_{\Phi,\Gamma} \quad & L_{ps}(\Phi,\Gamma) \\
\text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\
& p_k \geq 0, \quad \forall k \\
& \sum_{j=1}^{J} \gamma_{jk} = 1, \quad \forall k \\
& \gamma_{jk} \geq 0, \quad \forall j,k \\
& \gamma_{jk} = \tilde{z}^*_{jk}\gamma_{jk}, \quad \forall j,k \\
& \gamma_{jk} = \gamma_{jl}, \quad \text{if } \exists h \in \mathit{Child}(j,\mathcal{T}) \text{ such that } k,l \in \mathit{Leaf}(h,\mathcal{T}),
\end{aligned} \tag{4.56}$$

where

$$L_{ps}(\Phi,\Gamma) = \sum_{i=1}^{n}\sum_{j=1}^{J} \tau_{ij} \log \sum_{k=1}^{K} \gamma_{jk} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k).$$

By incorporating the second last constraint on Eq. (4.56) into the log-likelihood function, we define a HCML-specified log-likelihood function

$$\begin{aligned}
L_{\mathcal{T},hcml}(\Phi,\Gamma) \quad = \quad & \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_{ij} \log \gamma_{jj} p_j f_j(\mathbf{x}_i|\boldsymbol{\theta}_j) + \\
& \sum_{i=1}^{n}\sum_{j=K+1}^{J} \tau_{ij} \log \sum_{h\in \mathit{Child}(j,\mathcal{T})} \sum_{k\in \mathit{Leaf}(h,\mathcal{T})} \gamma_{jk} p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k).
\end{aligned} \tag{4.57}$$

Then an optimization problem Eq. (4.58) can be reformulated as

$$\begin{aligned}
\max_{\Phi,\Gamma} \quad & L_{\mathcal{T},hcml}(\Phi,\Gamma) \\
\text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\
& p_k \geq 0, \quad \forall k \\
& \sum_{j=1}^{J} \gamma_{jk} = 1, \quad \forall k \\
& \gamma_{jk} \geq 0, \quad \forall j,k, \\
& \gamma_{jk} = \gamma_{jl}, \quad \text{if } \exists h \in \mathit{Child}(j,\mathcal{T}) \text{ such that } k,l \in \mathit{Leaf}(h,\mathcal{T}).
\end{aligned} \tag{4.58}$$

To solve Eq. (4.58) we propose a new EM algorithm.

At starting the EM algorithm we initial an estimate of Φ and Γ. Particularly Γ is estimated by

$$\gamma_{jk}^{(1)} = \begin{cases} 0 & \text{if } j = 1, \cdots, J,\ k \notin \mathit{Leaf}(j, \mathcal{T}) \\ 1/\ \mathit{Depth}(\mathcal{T}) & \text{if } j = K+1, \cdots, J,\ k \in \mathit{Leaf}(j, \mathcal{T}) \\ 1 - \sum l = K+1^{J} \gamma_{lk}^{(1)} & \text{if } j = 1, \cdots, K,\ k = j \end{cases} \tag{4.59}$$

Let us have current estimates $\Phi^{(q)}$ and $\Gamma^{(q)}$ at starting the qth iteration of the EM algorithm. In E-step, a conditional expectation of z_{ik} is computed by

$$\begin{aligned} w_{ik}^{(q)} &= \mathbb{E}\left[z_{ik} \middle| \mathbf{x}_i, \tilde{\mathbf{z}}_i, \Phi^{(q)}, \Gamma^{(q)} \right] \\ &= \prod_{j \in Anc(k,\mathcal{T}) \cup \{k\}} \left[\frac{\gamma_{jk}^{(q)} p_k^{(q)} f_k(\mathbf{x}_i | \boldsymbol{\theta}_k^{(q)})}{\sum_{l \in Leaf(j,\mathcal{T})} \gamma_{jl}^{(q)} p_l^{(q)} f_l(\mathbf{x}_i | \boldsymbol{\theta}_l^{(q)})} \right]^{\tau_{ij}}. \end{aligned} \tag{4.60}$$

With the conditional expectation, a conditional log-likelihood function is defined by

$$\begin{aligned} Q_{\mathcal{T},hcml}(\Phi, \Gamma | \Phi^{(q)}, \Gamma^{(q)}) &= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*} \gamma_{jk} p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} w_{ik}^{(q)} \log \tilde{z}_{jk}^{*} \gamma_{jk} + \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) \\ &= Q_{\mathcal{T}}(\Gamma | \Gamma^{(q)}) + Q(\Phi | \Phi^{(q)}), \end{aligned} \tag{4.61}$$

where $Q(\Phi|\Phi^{(q)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_{ij} \sum_{k=1}^{K} w_{ik}^{(q)} \log p_k f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is defined in Eq. (2.4) and $Q_{\mathcal{T}}(\Gamma|\Gamma^{(q)})$ is defined by

$$Q_{\mathcal{T}}(\Gamma | \Gamma^{(q)}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{K} \tau_{ij} \log \gamma_{jj} + \sum_{j=K+1}^{J} \tau_{ij} \sum_{k \in Leaf(j,\mathcal{T})} w_{ik}^{(q)} \log \gamma_{jk} \right]. \tag{4.62}$$

Sets of constraints on Φ and Γ in Eq. (4.58) are independent from each other. The M-step of maximizing a conditional log-likelihood function is therefore composed of two parts: maximization of $Q(\Phi|\Phi^{(q)})$ and maximization of $Q_{\mathcal{T}}(\Gamma|\Gamma^{(q)})$.

First let $\Phi^{(q+1)}$ be a solution to the following maximization problem:

$$\begin{aligned} \max_{\Phi} \quad & Q(\Phi | \Phi^{(q)}) \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0, \quad \forall k. \end{aligned}$$

Solving the optimization problem is identical to the M-step in unsupervised learning that is described in Section 2.1.2:

$$\begin{aligned} p_k^{(q+1)} &= \sum_{i=1}^{n} w_{ik}^{(q)} / n, \\ \boldsymbol{\theta}_k^{(q+1)} &= \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k). \end{aligned}$$

In addition we obtain $\Gamma^{(q+1)}$ by solving the following optimization problem:

$$\begin{aligned} \max_{\Gamma} \quad & Q(\Gamma|\Gamma^{(q)}) \\ \text{s.t.} \quad & \sum_{j=1}^{J} \gamma_{jk} = 1, \quad \forall k \\ & \gamma_{jk} \geq 0, \quad \forall j,k, \\ & \gamma_{jk} = \gamma_{jl}, \quad \text{if } \exists h \in \mathit{Child}(j,\mathcal{T}) \text{ such that } k,l \in \mathit{Leaf}(h,\mathcal{T}). \end{aligned} \tag{4.63}$$

Because of difficulties in finding a closed form optimal solution to Eq. (4.63), we can use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to obtain $\Gamma^{(q+1)}$ as done under PCML in Section 4.2.3.

A detailed EM algorithm for learning finite mixture models on AVT under HCML is described in Figure 4.5.

4.5 Model Selection

It is noteworthy that PCML is feasible a feasible model of missing label mechanisms for partially supervised learning on AVT. In Section 3.4.1 we have discussed that we always find a feasible estimate of FMM if all the precise labels are observable or if unlabeled data is observable. On AVT all the precise labels are represented by the leaf nodes, while unlabeled data is corresponding to the root node. We therefore do not need concern feasibility when we find MLEs of FMM under PCML. In fact Eq. (4.59) that has been used to generate initial estimates of γ_{jk} under HCML also represents feasible estimates of γ_{jk} for PCML.

A solution space under HCML includes PCML solution space. In Eq. (4.48) HCML parameterizes ξ_{jl} which is completely determined by the other parameters under PCML. Therefore the maximum log-likelihood under HCML is obviously as large as the maximum log-likelihood

1: $q \leftarrow 0$

2: Initialize $\gamma_{jk}^{(1)}$ by Eq. (4.59).

3: Initialize $\Phi^{(1)}$

4: **repeat**

5: $\quad q \leftarrow q+1$

6: $\quad$ E-step: $w_{ik}^{(q)} \leftarrow \sum_{j=1}^{J} tau_{ij} \left[\gamma_{jk}^{(q)} p_k^{(q)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q)}) / \sum_{l \in Leaf(j,\mathcal{T})} \gamma_{jl}^{(q)} p_l^{(q)} f_l(\mathbf{x}_i|\boldsymbol{\theta}_l^{(q)})\right]$

7: $\quad$ M-step(1): $p_k^{(q+1)} \leftarrow \sum_{i=1}^{k} w_{ik}^{(q)}/n$

8: $\quad$ M-step(2): $\boldsymbol{\theta}_k^{(q+1)} \leftarrow \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} w_{ik}^{(q)} \log f_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$

9: $\quad$ M-step(3): Obtain $\gamma_{jk}^{(q+1)}$ by solving Eq. (4.63) using BFGS method.

10: **until** $L_{\mathcal{T},hcml}(\Phi^{(q+1)}, \Gamma^{(q+1)}) - L_{ps,pcml}(\Phi^{(q)}, \Gamma^{(q)}) < \epsilon$

11: **return** $\hat{\Phi} \leftarrow \Phi^{(q)}$, $\hat{\Gamma} \leftarrow \Gamma^{(q)}$

Figure 4.5: EM-HCML: EM algorithm for learning FMM under HCML mechanism on AVT

under PCML. In addition CPCML generalizes HCML by relaxing constraint Eq. (4.55). We therefore find that

$$L_{ps,pcml}(\hat{\Phi}, \hat{\boldsymbol{\gamma}}) \leq L_{\mathcal{T},hcml}(\hat{\Phi}, \hat{\Gamma}) \leq L_{ps,cpcml}(\hat{\Phi}, \hat{\Gamma})$$

on AVT $\mathcal{T}$, where the MLE of estimates are the results from EM-PCML, EM-HCML, and EM-CPCML, respectively.

We again use AIC criteria to select the most plausible missing label mechanism on the observed data set $\mathcal{D}_{obs}$. Let $|\Phi|$ be the number of parameters in Φ to be used in computing AIC. Then $|\Phi|$ is the same under PCML, HCML, and CPCML. In addition to $|\Phi|$, the number of parameters in Γ affects AIC values under HCML. Now we recall another parameterization of FMM in Section 4.4.1 that has used Ξ rather than Γ to represent FMM. Because of the equality constraint

$$\sum_{l \in Child(j,\mathcal{T})} \xi_{jl} = 1, \quad j = K+1, \cdots, J,$$

that has been defined in Eq. (4.49), we have that the number of parameters on Ξ is

$$\sum_{j=K+1}^{J} (|\, Child(j,\mathcal{T})| - 1). \tag{4.64}$$

Therefore AIC under HCML is defined by

$$AIC_{\mathcal{T},hcml} = -2L_{\mathcal{T},pcml}(\hat{\Phi}, \hat{\Gamma}) + 2|\Phi| + 2 \sum_{j=K+1}^{J} (|\, Child(j, \mathcal{T})| - 1), \tag{4.65}$$

where $|\, Child(j, \mathcal{T})|$ represents the number of child nodes of j on $\mathcal{T}$. We define that the best or the most plausible missing label mechanism on $\mathcal{D}_{obs}$ on AVT is corresponding to the least AIC value among $AIC_{ps,pcml}$, $AIC_{\mathcal{T},hcml}$, and $AIC_{ps,cpcml}$.

4.6 Summary

In this chapter we have newly parameterized FMM and specified previously proposed EM algorithms in Chapter 3 to cases that observable labels are defined by attribute value taxonomies. This chapter delivered an intuitive concept of partially supervised FMM learning that the overall FMM is composed of component FMMs each of which is corresponding to each specific partial label. Partially supervised FMM learning therefore tries to estimate a mixture of mixtures. Such insight led to newly defined missing label mechanism, which is named hierarchy-conditional missing label mechanism (HCML). We proposed EM algorithm of estimating FMM under HCML (EM-HCML) that provides flexibilities compared to EM-PCML and specificities compared to EM-CPCML. We suggested AIC to be used for selecting the most plausible missing label mechanism underlying on the observed data.

In the following chapter we evaluate the performance of EM-PCML, EM-HCML, and EM-CPCML on synthetic data on exponential survival trees by comparing with supervised, unsupervised, and semi-supervised FMM learning. We also conduct a case study of exponential survival time modeling for gastric cancer patients from which this research has originally been motivated.

CHAPTER 5. EXPERIMENTAL RESULTS

In this chapter we show how the proposed EM algorithms for partially supervised learning perform compared to supervised learning, unsupervised, and semi-supervised learning. First we show how the MLEs obtained by the proposed algorithms are close to the true exponential survival tree model that has been introduced by Davis and Anderson (1989). Second we apply the proposed EM algorithms to estimate survival time models for patients with the gastric signet ring cell carcinoma on the cardia on SEER research data (National Cancer Institute, 2011).

5.1 Simulations on Exponential Survival Tree

In this section we show the performance and the advantages of partially supervised learning algorithms on AVT-guided survival time data by using a synthetic data set introduced by Davis and Anderson (1989).

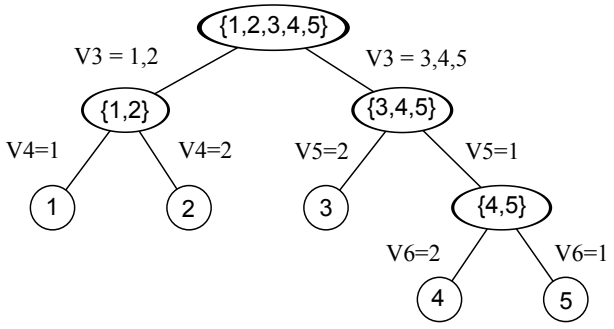


Figure 5.1: An exponential survival tree simulated in Davis and Anderson (1989).

Table 5.1: Parameters of the mixture of exponential survival time distributions in Davis and Anderson (1989)

Class k	Proportion p_k	Death rate λ_k
1	0.20	0.35
2	0.20	0.60
3	0.30	0.80
4	0.15	1.00
5	0.15	1.75

5.1.1 Data description

Davis and Anderson (1989) studied an classification tree algorithm to split heterogeneous population into subgroups that have a homogeneous exponential survival time distribution within each subgroup. In their simulation studies they have defined a data set with five classes which are hierarchically specified by using four variables (Figure 5.1). $V3$ is an ordinary variable that take integer values from one to five with equal chances. The other three variables, $V4$, $V5$ and $V6$, are binary variables which take value one or two with probability $1/2$ in each. Each data i has been categorized into one of five classes by the following classification rules after observing variables from $V3$ to $V6$:

- Class 1 if $V3 \in \{1, 2\}$ and $V4 = 1$
- Class 2 if $V3 \in \{1, 2\}$ and $V4 = 2$
- Class 3 if $V3 \in \{3, 4, 5\}$ and $V5 = 2$
- Class 4 if $V3 \in \{3, 4, 5\}$, $V5 = 1$ and $V6 = 2$
- Class 5 if $V3 \in \{3, 4, 5\}$, $V5 = 1$ and $V6 = 1$

Let λ_k be a constant death rate within class k, which has been set to be

$$\lambda_1 = 0.35, \lambda_2 = 0.60, \lambda_3 = 0.80, \lambda_4 = 1.00, \lambda_5 = 1.75.$$

As a result a mixture of exponential survival time distributions is defined with parameters in Table 5.1.

1: Define right-censoring time distribution $f_{T_c}(t)$ and coarsening probabilities Γ
2: **for** $i = 1$ **to** n **do**
3: Generate $V3$, $V4$, $V5$, $V6$
4: Set class k by following the predefined classification rules, $z_{ik} \leftarrow 1$
5: Generate survival time T from $f_T(t) = \lambda_k e^{-\lambda_k t}$
6: Generate right-censoring time T_c from $f_{T_c}(t)$
7: $t_i \leftarrow \min\{T, T_c\}$
8: $c_i \leftarrow \mathbb{I}(T \leq T_c)$
9: **end for**

Figure 5.2: Generation of synthetic data set based on Davis and Anderson (1989)

Time to death T within class k is distributed by

$$P(T > t|\text{class } k) = e^{-\lambda_k t}.$$

Davis and Anderson (1989) conducted separate simulations by using two right-censoring time distribution: Uniform(2,4) and Uniform(0.25,1.25). After generating survival time T and censoring time T_c for each instance, the earlier time between T and T_c is stored with indicating the stored time is whether actual failure time or not. With the first censoring time distribution about 85% instances are supposed to be observed with their actual survival time so that $t_i = T$ and $c_i = 1$ while 15% are observed with right-censoring time so that $t_i = T_c$ and $c_i = 0$. On the other hand with the second censoring time distribution actual survival time for only 43% instances is observed so that the ambiguity of data is increased from the first censoring time distribution. A detail of data generation is described in Figure 5.2.

A difference of this study from Davis and Anderson (1989) is that we allow missing data to appear. By assuming the classification of instances is performed by sequentially observing V3 to V6 based on splitting criteria in Figure 5.1, we define nine observable class labels as shown in Table 5.2. After generating synthetic data by using the algorithm in Figure 5.2 for each data i we generate a coarsened version of class label $\tilde{\mathbf{z}}_i$ by coarsening probability γ_{jk} such that

$$\gamma_{jk} = P(\tilde{\mathbf{z}}_i = \tilde{\mathbf{z}}_j^* | z_{ik} = 1).$$

Pattern j	Class k				
	1	2	3	4	5
1	.8				
2		.8			
3			.8		
4				.7	
5					.7
6	.1	.1			
7				.1	.1
8			.1	.1	.1
9	.1	.1	.1	.1	.1

(a) PCML1

Pattern j	Class k				
	1	2	3	4	5
1	.6				
2		.6			
3			.6		
4				.3	
5					.3
6	.3	.3			
7				.3	.3
8			.3	.3	.3
9	.1	.1	.1	.1	.1

(b) PCML2

Pattern j	Class k				
	1	2	3	4	5
1	.4				
2		.4			
3			.4		
4				.1	
5					.1
6	.3	.3			
7				.3	.3
8			.3	.3	.3
9	.3	.3	.3	.3	.3

(c) PCML3

Pattern j	Class k				
	1	2	3	4	5
1	.6				
2		.2			
3			.4		
4				.7	
5					.3
6	.1	.5			
7				.1	.5
8			.5	.1	.1
9	.3	.3	.1	.1	.1

(d) HCML1

Pattern j	Class k				
	1	2	3	4	5
1	.2				
2		.6			
3			.4		
4				.3	
5					.7
6	.5	.1			
7				.5	.1
8			.5	.1	.1
9	.3	.3	.1	.1	.1

(e) HCML2

Pattern j	Class k				
	1	2	3	4	5
1	.4				
2		.4			
3			.2		
4				.3	
5					.3
6	.1	.5			
7				.1	.5
8			.3	.5	.1
9	.5	.1	.5	.1	.1

(f) CPCML1

Pattern j	Class k				
	1	2	3	4	5
1	.4				
2		.4			
3			.2		
4				.3	
5					.3
6	.5	.1			
7				.5	.1
8			.3	.1	.5
9	.1	.5	.5	.1	.1

(g) CPCML2

Pattern j	Class k				
	1	2	3	4	5
1	.4				
2		.1			
3			.4		
4				.3	
5					.1
6	.5	.1			
7				.5	.1
8			.3	.1	.5
9	.1	.8	.5	.1	.3

(h) CPCML3

Figure 5.3: Eight sets of coarsening probabilities γ_{jk}

Table 5.2: Observable class label patterns based on the classification tree in Davis and Anderson (1989). '-' represents missing data.

Pattern	Classification variables				Coarsened label $\tilde{\mathbf{z}}_j^*$				
j	V3	V4	V5	V6	$\tilde{z}_{j1}^*$	$\tilde{z}_{j2}^*$	$\tilde{z}_{j3}^*$	$\tilde{z}_{j4}^*$	$\tilde{z}_{j5}^*$
1	{1,2}	1	-	-	1	0	0	0	0
2	{1,2}	2	-	-	0	1	0	0	0
3	{3,4,5}	-	2	-	0	0	1	0	0
4	{3,4,5}	-	1	2	0	0	0	1	0
5	{3,4,5}	-	1	1	0	0	0	0	1
6	{1,2}	-	-	-	1	1	0	0	0
7	{3,4,5}	-	1	-	0	0	0	1	1
8	{3,4,5}	-	-	-	0	0	1	1	1
9	-	-	-	-	1	1	1	1	1

We use eight different sets of γ_{jk} that three sets represent PCML, two sets represent HCML, and the other three sets represent CPCML missing label mechanisms (Figure 5.3). It must be noted that OCML is impossible to be defined with this simulation. To show how the sample size affects the performance of the estimation, we used three levels of sample sizes n: 500, 2000 and 8000. We randomly generated 100 sample data sets for each combination of two censoring time distributions, eight missing label mechanisms and six sample sizes. So we have performed the estimation of finite mixture models on 4800 data sets in total. For each random data set we estimated finite mixture models by using seven learning algorithms: supervised, unsupervised, semi-supervised under CML (Semi-CML), semi-supervised under CCML (Semi-CCML), and three partially supervised learning methods (EM-PCML, EM-HCML and EM-CPCML). The first four methods from existing studies have been described in Chapter 2, while the latter three have been proposed in Chapters 3–4. Therefore we obtain 100 estimates of a mixture of exponential survival time distributions under each combination of simulation parameter values in Table 5.3. The 100 estimates are capable for pairwise comparisons across learning algorithms because all the seven algorithms shared the data sets for each combination of the other three simulation parameters.

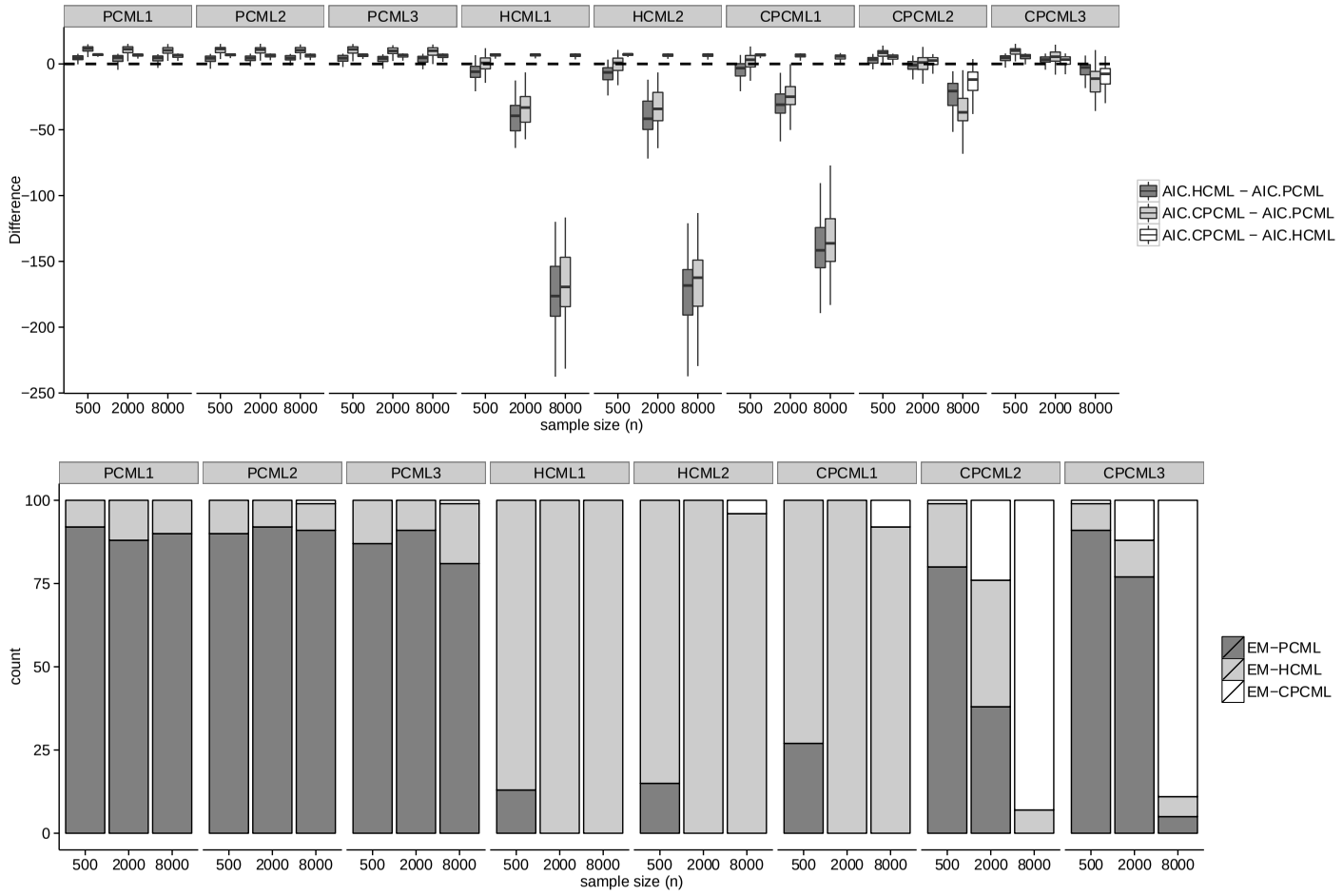


Figure 5.4: Comparison of partially supervised estimates based on AIC values on data with U(2,4) censoring time distribution

Table 5.3: Simulation parameters for estimating the mixture of five exponential survival time distributions in Davis and Anderson (1989)

Parameter	Level
Censoring time distribution	U(2,4), U(0.25,1.25)
Missing label mechanism	PCML1, PCML2, PCML3, HCML1, HCML2, CPCML1, CPCML2, CPCML3
Sample size	500, 2000, 8000
Learning algorithm	Supervised, Unsupervised, Semi-CML, Semi-CCML, EM-PCML, EM-HCML, EM-CPCML

5.1.2 Comparison between partially supervised learning algorithms

Figure 5.4 shows model selection criteria with U(2,4) censoring time distribution. For PCML1, PCML2 and PCML3, the estimates from EM-PCML were consistently preferred to the estimates from EM-HCML or EM-CPCML as they are supposed to be. Because EM-PCML involves all the assumptions on PCML mechanisms with the least number of parameters, the estimates from EM-PCML on data sets for PCML1, PCML2, and PCML3 are expected to be unbiased and more robust than the results from EM-HCML and EM-CPCML. Regardless the sample size EM-PCML has shown the least AIC values in over than 80% cases. For HCML1 and HCML2 on the other hand EM-HCML was consistently preferred to EM-PCML. For data sets with HCML mechanisms EM-CPCML was also superior to EM-PCML for large sample sizes ($n = 2000, 8000$), while EM-PCML was still preferred to EM-CPCML with small samples ($n = 500$). Even with large sample sizes however AIC values from EM-CPCML were slightly greater than AIC values from EM-HCML. With three data sets under CPCML1, CPCML2 and CPCML3 that violate HCML mechanisms, the preferences of estimates depended on data sets. Under CPCML1 the preferences of estimates were similar to those under HCML mechanisms, so EM-HCML was most preferred regardless sample sizes. On the other side under CPCML2 and CPCML3 EM-CPCML was most preferred with very large sample sizes ($n = 8000$), while estimates with less parameters from EM-PCML or EM-HCML led to less AIC values for smaller sample sizes ($n = 500, 2000$).

MLEs of the FMM that were obtained from the three partially supervised EM algorithms under U(2,4) censoring time are shown in Figure 5.5. On data sets generated under PCML

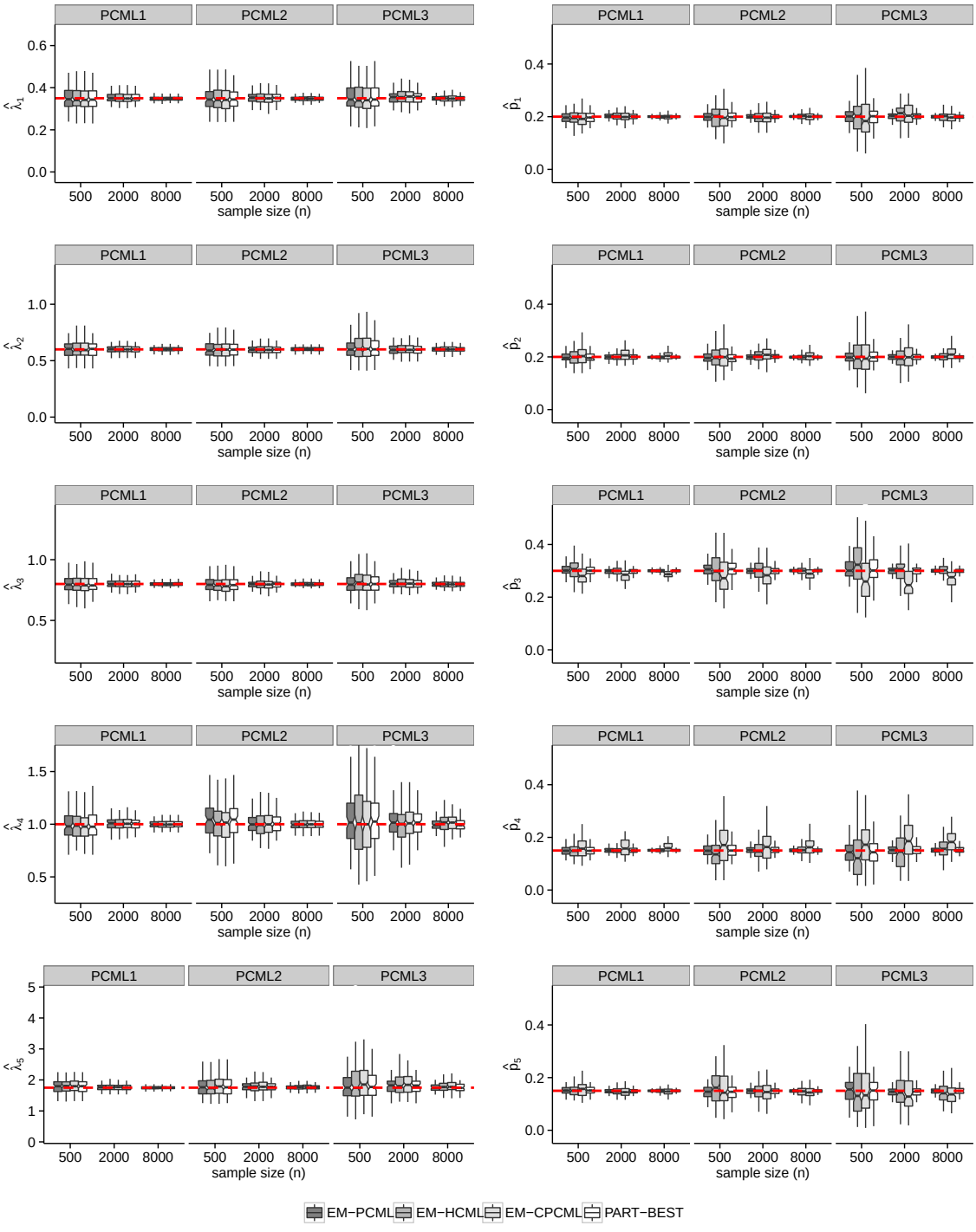


Figure 5.5: MLE of FMM from partially supervised learning on synthetic data sets with U(2,4) censoring time

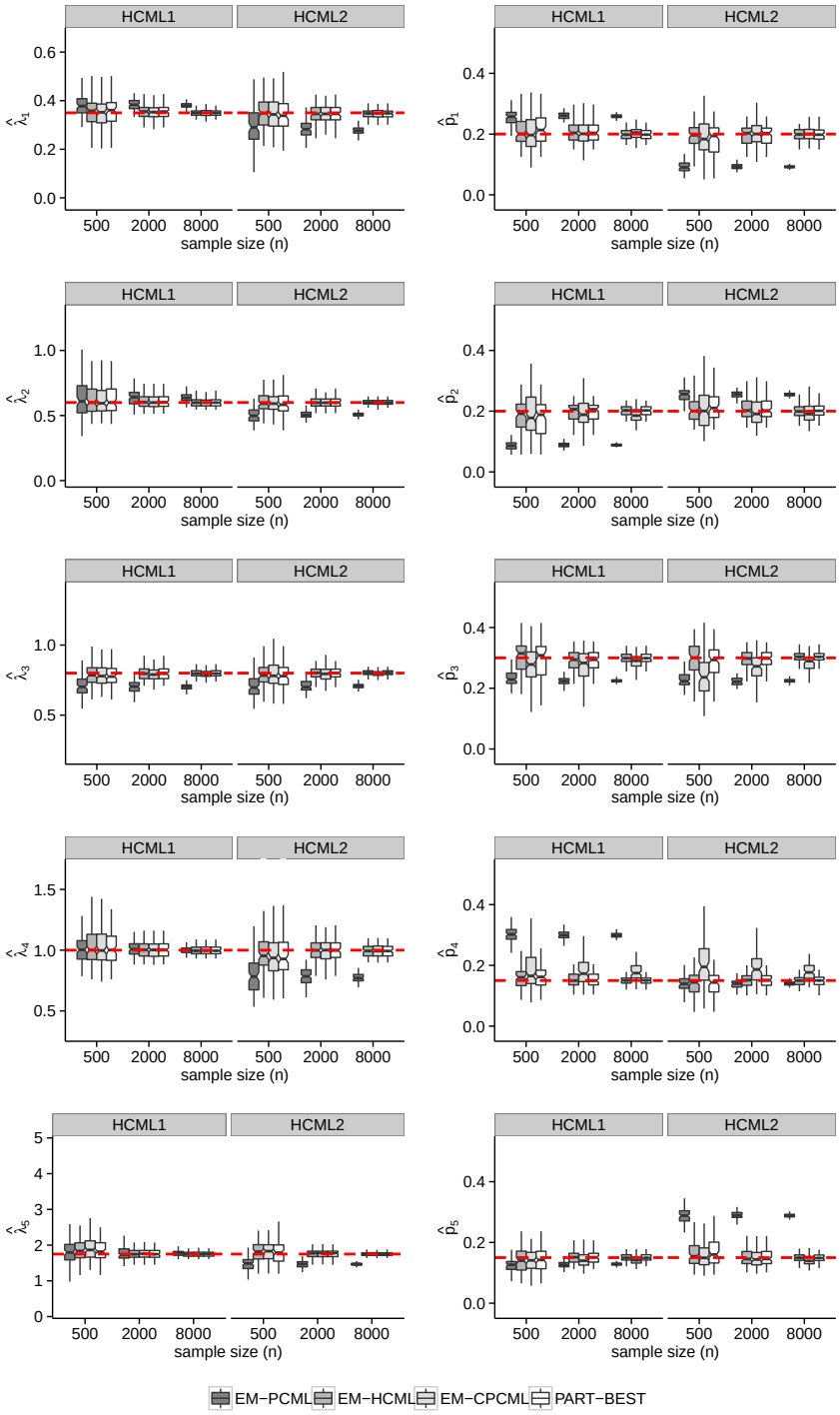


Figure 5.5: (Continued)

MLE of FMM from partially supervised learning on synthetic data sets with U(2,4) censoring time

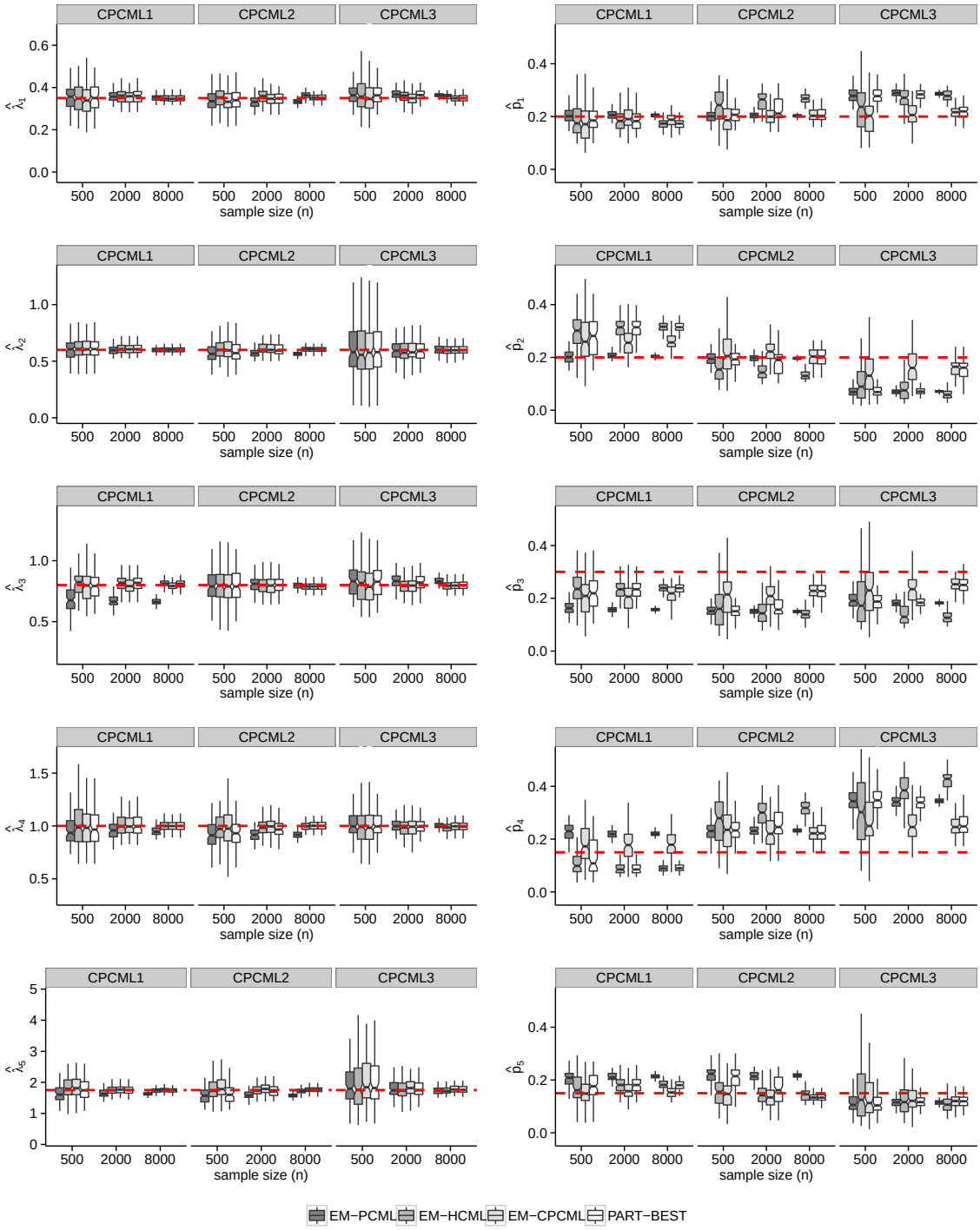


Figure 5.5: (Continued)

MLE of FMM from partially supervised learning on synthetic data sets with U(2,4) censoring time

mechanisms EM-PCML and EM-HCML consistently led to unbiased estimates, while EM-CPCML tended to underestimate p_3 and overestimate p_4. In addition the variances of estimates from EM-PCML were smaller than those from EM-HCML and EM-CPCML. We therefore consider EM-PCML be the best partially supervised learning algorithm for data under PCML mechanism. Under HCML mechanisms on the other hand EM-PCML produced biased estimates because PCML assumption was violated. EM-CPCML still produced biased estimates of p_3 and p_4, which looks caused by insufficient numbers of samples. By looking $\hat{p}_3$ and $\hat{p}_4$ we can see that the estimates from EM-CPCML are getting closer to the true parameter values as sample sizes increase. With a small number of samples however EM-HCML produced unbiased estimates. In simulations under HCML mechanisms therefore EM-HCML was considered the best learning algorithm. Under three CPCML mechanisms we have found that all the partially supervised learning algorithms produced biased estimates. EM-PCML and EM-HCML were supposed to obtain biased estimates because PCML and HCML assumptions were violated on the CPCML data sets. The reason of biased estimates from EM-CPCML is that sample sizes were not enough to obtain unbiased estimates by using a large number of parameters. So we could observe that bias of estimates obtained by EM-CPCML has been decreased as sample sizes increase, while bias of estimates obtained by EM-PCML and EM-HCML has been consistent or even increased. For a large sample size ($n = 8000$) therefore EM-CPCML looked superior to EM-PCML and EM-HCML. Such observations agree with model selections based on the AIC criteria that have been shown in Figure 5.4. In Figure 5.5 'PART-BEST' represents the MLE with the least AIC values among MLEs from the three algorithms for each random data set. So we consider PART-BEST as the best estimates we obtained by using partially supervised learning.

With U(0.25,1.25) censoring time distribution model selections based on AIC values similarly performed to the previous simulation under PCML. Under HCML mechanisms EM-PCML tended to be more frequently selected with a small sample size ($n = 500$) than U(2,4) censoring time cases. EM-HCML however still dominated to the other learning methods under all the HCML mechanisms. On the other side under CPCML mechanisms EM-CPCML has not been preferred to EM-PCML or EM-HCML even with a large sample size ($n = 8000$). As sam-

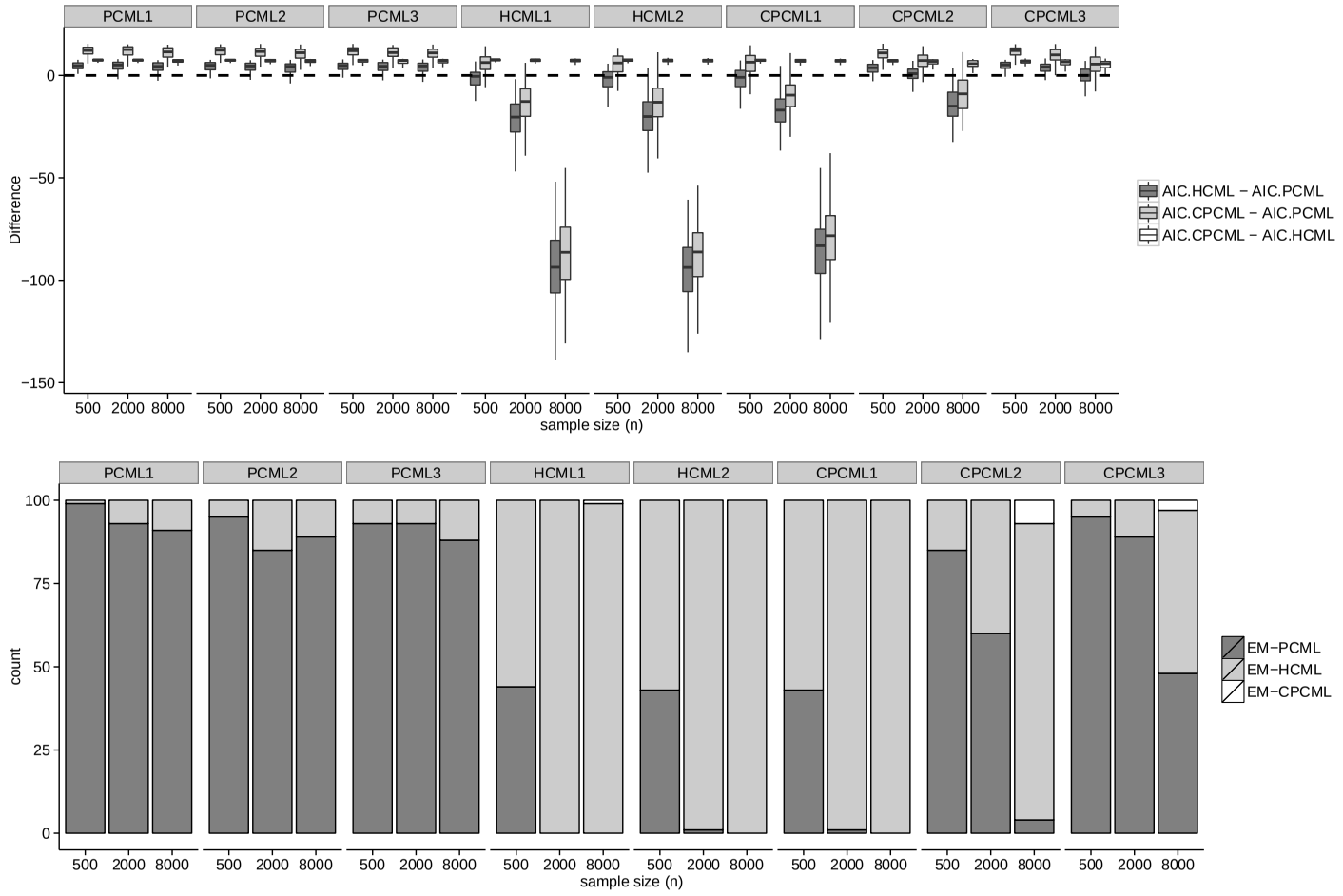


Figure 5.6: Comparison of partially supervised estimates based on AIC values on data with U(0.25,1.25) censoring time distribution

ple sizes increase EM-HCML has been preferred to EM-PCML on data sets generated under CPCML mechanisms because EM-HCML allowed more flexibilities to estimate missing label mechanisms than EM-PCML. Although EM-CPCML used the most appropriate models on data sets under CPCML mechanisms, observed data sets contained too many ambiguities to obtain robust estimates by using EM-CPCML algorithm. While 85% of actual survival time is observed with U(2,4) censoring time distribution, only 43% of actual survival time is observed with U(0.25,1.25) censoring time. Such reduced chances of observing actual survival time require more sample sizes to obtain estimates as informative as U(2,4) cases.

MLEs of λ_k in Figure 5.7 obtained from data sets with U(0.25,1.25) have larger variances than Figure 5.5. It shows that increased ambiguities in observed survival time data led to increased uncertainties in estimating survival time distributions. In addition MLEs of p_k from EM-CPCML in Figure 5.7 were far from the true parameter values compared to the results in Figure 5.5 under CPCML mechanisms. Although EM-CPCML looks superior to EM-PCML and EM-HCML under CPCML mechanisms in Figure 5.7, AIC-based model selection approaches have been failed to select EM-CPCML as the best learning method because increased likelihoods of observed data were not sufficiently large to make increased complexities of statistical models beneficial.

5.1.3 Comparison of partially supervised learning to conventional learning methods

In the previous section we compared the results from three different partially supervised learning methods. In this section we compare the best estimates from partially supervised learning methods to the estimates from supervised, unsupervised, and semi-supervised learning methods. Several issues in estimating finite mixture models by using the competitive methods are described below.

- **Supervised**: To estimate a five-component mixture by supervised learning methods we must observe at least one precisely labeled data from each of five classes. Such limitation occasionally made supervised learning methods fail to estimate the finite mixture models.

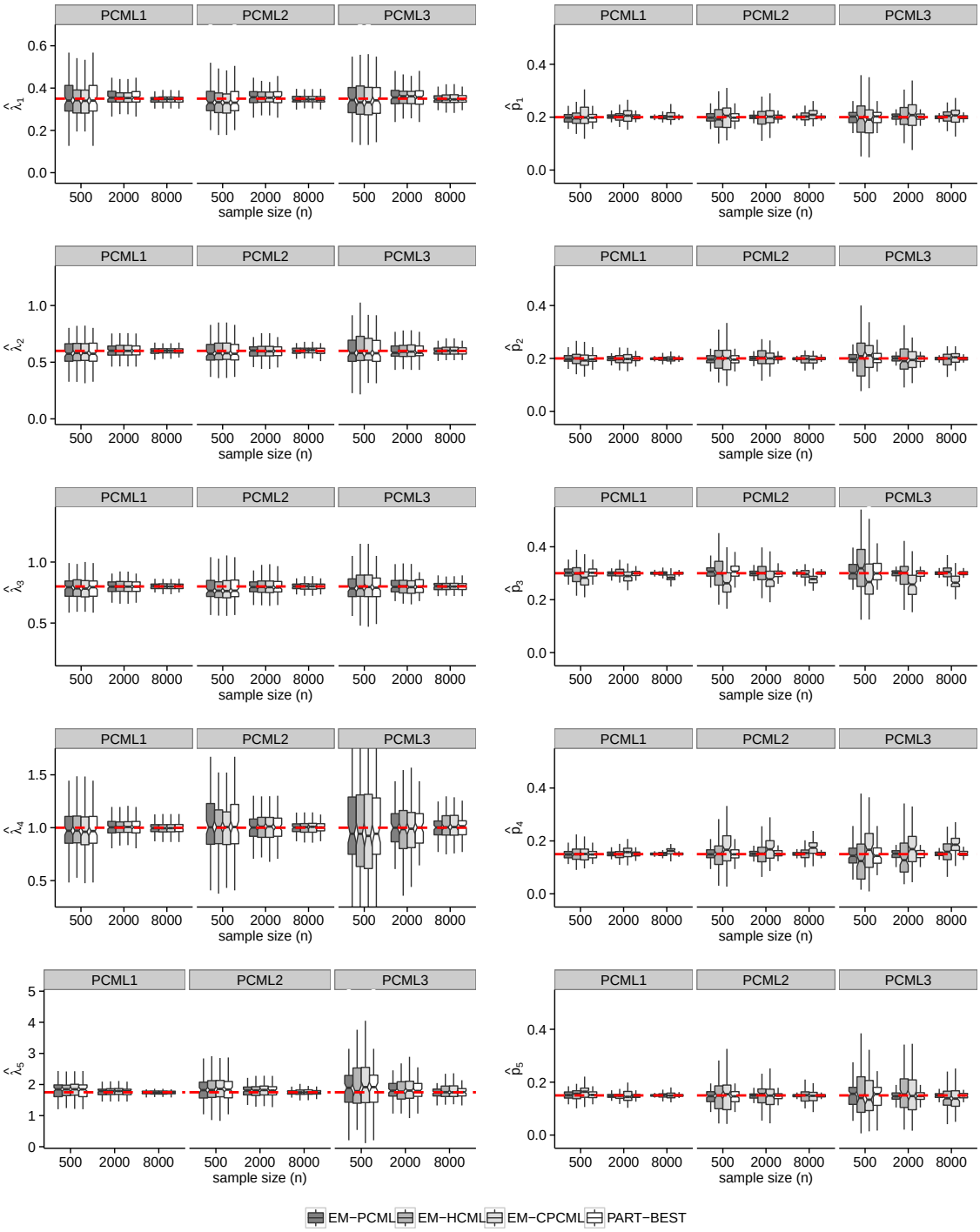


Figure 5.7: MLE of FMM from partially supervised learning on synthetic data sets with U(0.25,1.25) censoring time

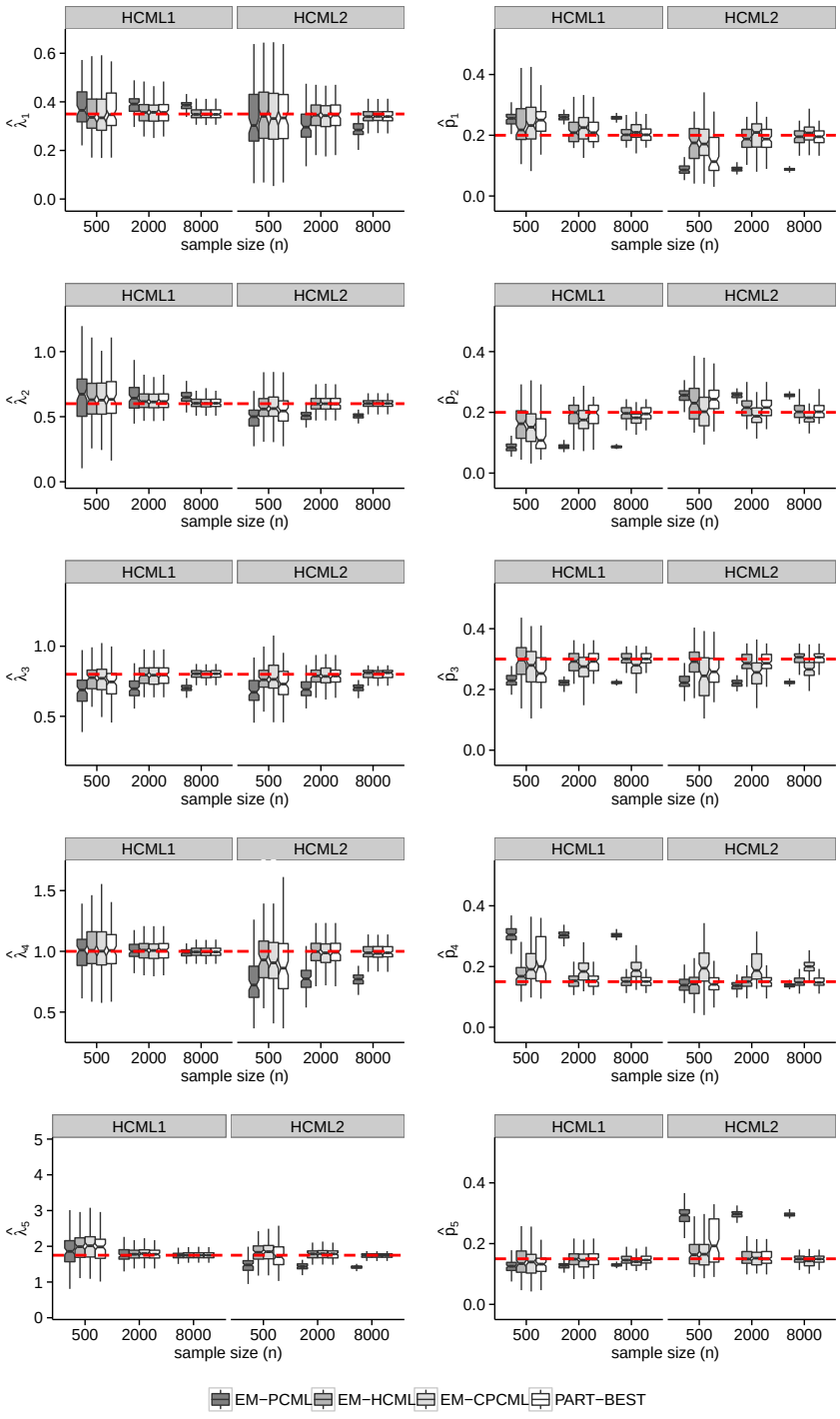


Figure 5.7: (Continued)

MLE of FMM from partially supervised learning on synthetic data sets with U(0.25,1.25) censoring time

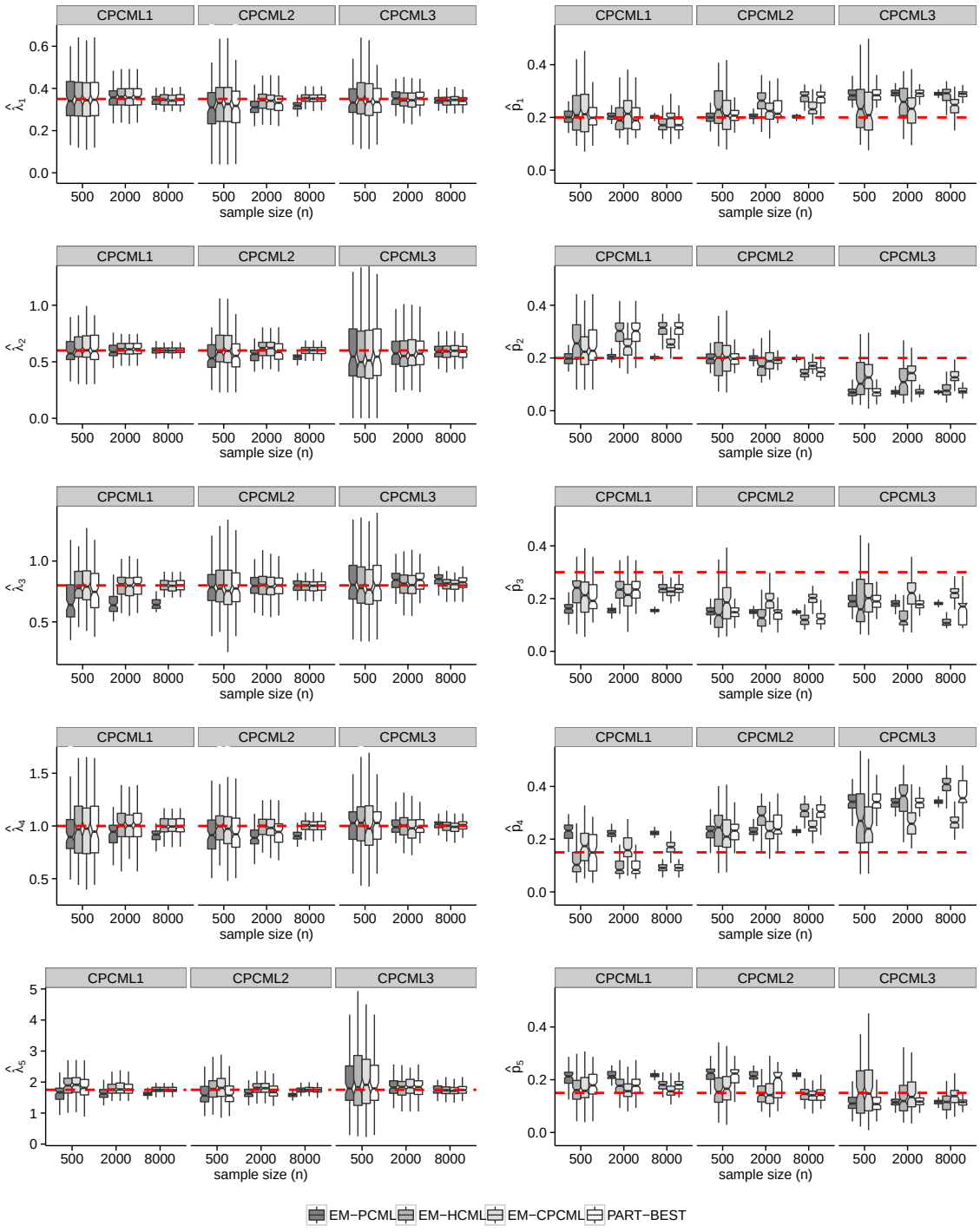
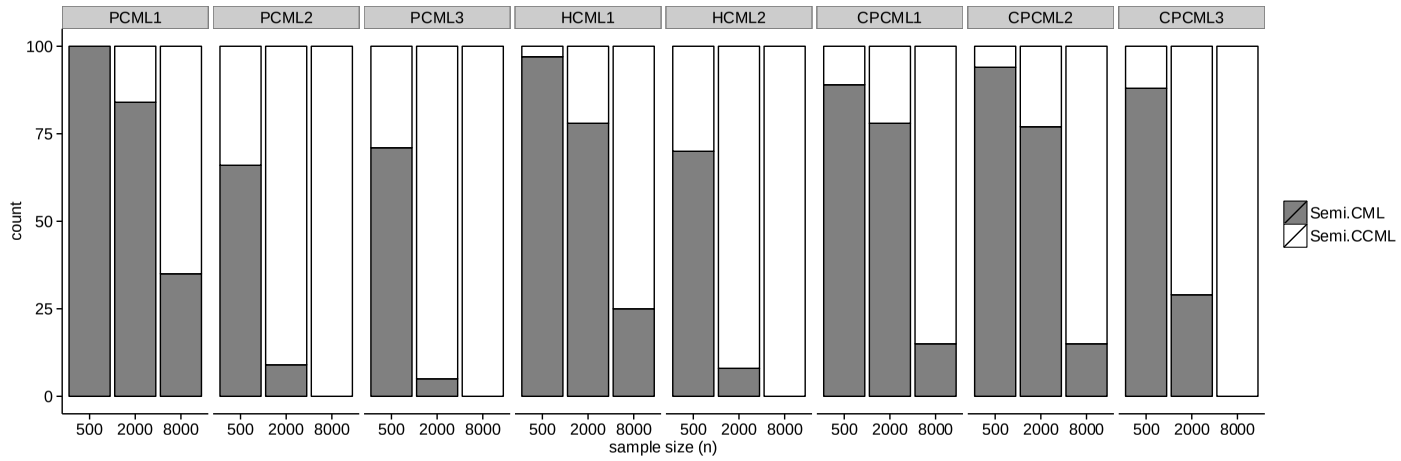
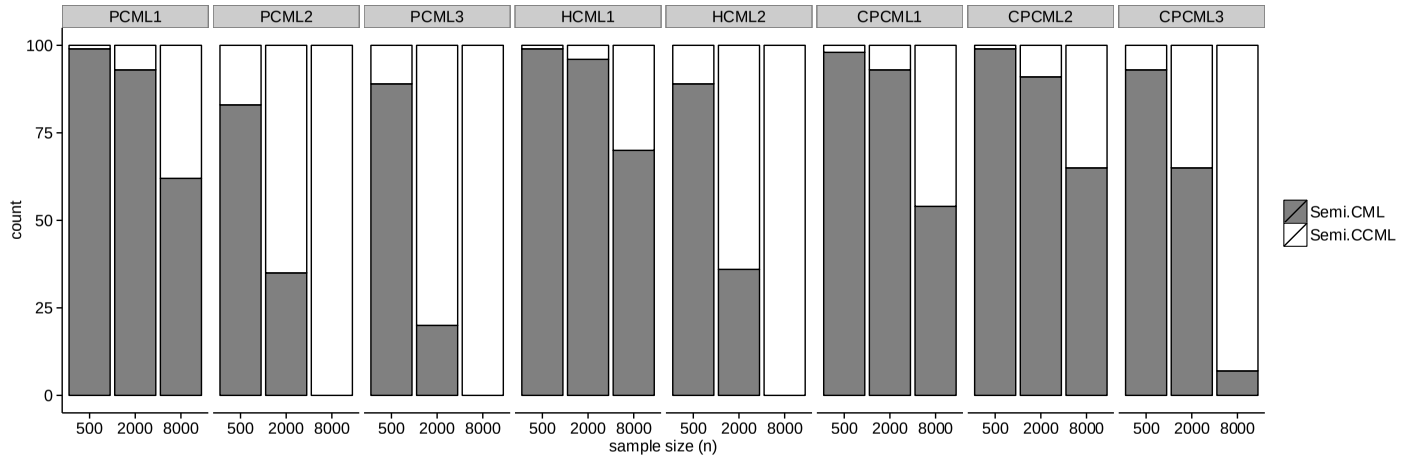


Figure 5.7: (Continued)

MLE of FMM from partially supervised learning on synthetic data sets with U(0.25,1.25) censoring time



(a) U(2,4)



(b) U(0.25,1.25)

Figure 5.8: Comparison of semi supervised estimates based on AIC values on synthetic data with (a) U(2,4) and (b) U(0.25,1.25) censoring time distributions

We drop the cases in evaluating the performance of supervised learning, while estimates on the same data sets from unsupervised, semi-supervised, and partially supervised learning are still reported.

- **Unsupervised**: Unsupervised learning does not use the order of class labels. It possibly causes mismatching of estimates $\hat{\lambda_k}$ and $\hat{p_k}$ to the parameter λ_k and p_k for each k. In this simulation we constrain the order of $\hat{\lambda}_k$'s to be

$$\hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \hat{\lambda}_3 \leq \hat{\lambda}_4 \leq \hat{\lambda}_5$$

in unsupervised learning cases. Because it includes additional valid information that has not been used in the other types of learning, we may overrate the performance of unsupervised learning compared to the others. Even with such advantages for unsupervised learning we will show partially supervised learning outperforms.

- **Semi-supervised** In Chapter 2 we introduced two types of semi-supervised learning methods: semi-supervised under CML mechanisms (Semi-CML) and semi-supervised under CCML mechanisms (Semi-CCML). We particularly found the best estimates in semi-supervised learning in the same way that we decided the best results in partially supervised learning based on AIC criteria (Figure 5.8). All the eight missing label mechanisms in Figure 5.3 violate CML mechanism that Semi-CML assumes. Therefore for large sample sizes ($n = 8000$) Semi-CCML has dominated Semi-CML. However for a small sample size ($n = 500$) the results from Semi-CML were more frequently selected as the better ones. We denote the better estimates in semi-supervised learning by Semi-BEST.

Figures 5.9 and 5.10 describe the MLEs obtained from supervised, unsupervised, semi-supervised, and partially supervised learning on 100 random data sets with U(2,4) and U(0.25, 1.25) censoring time distributions, respectively. We could find some remarkable advantages of using partial labels from the results.

For data sets under PCML mechanisms, distributions of MLEs from partially supervised learning (PART-BEST) were centered to the true parameter values regardless sample sizes, while the other learning methods produced biased estimates. Variances of $\hat{\lambda}_k$ from partially

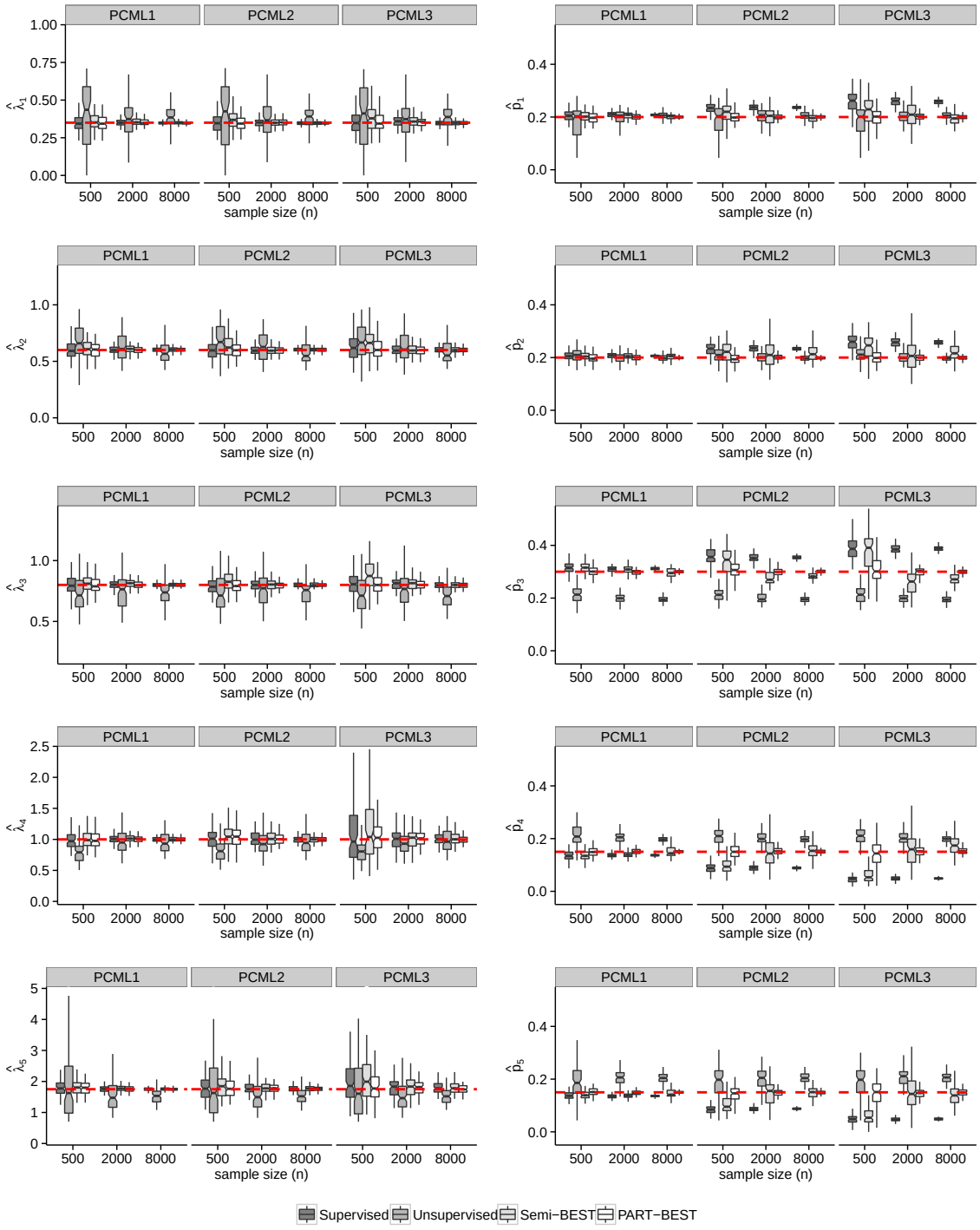


Figure 5.9: Comparison of MLEs on synthetic data sets with U(2,4) censoring time

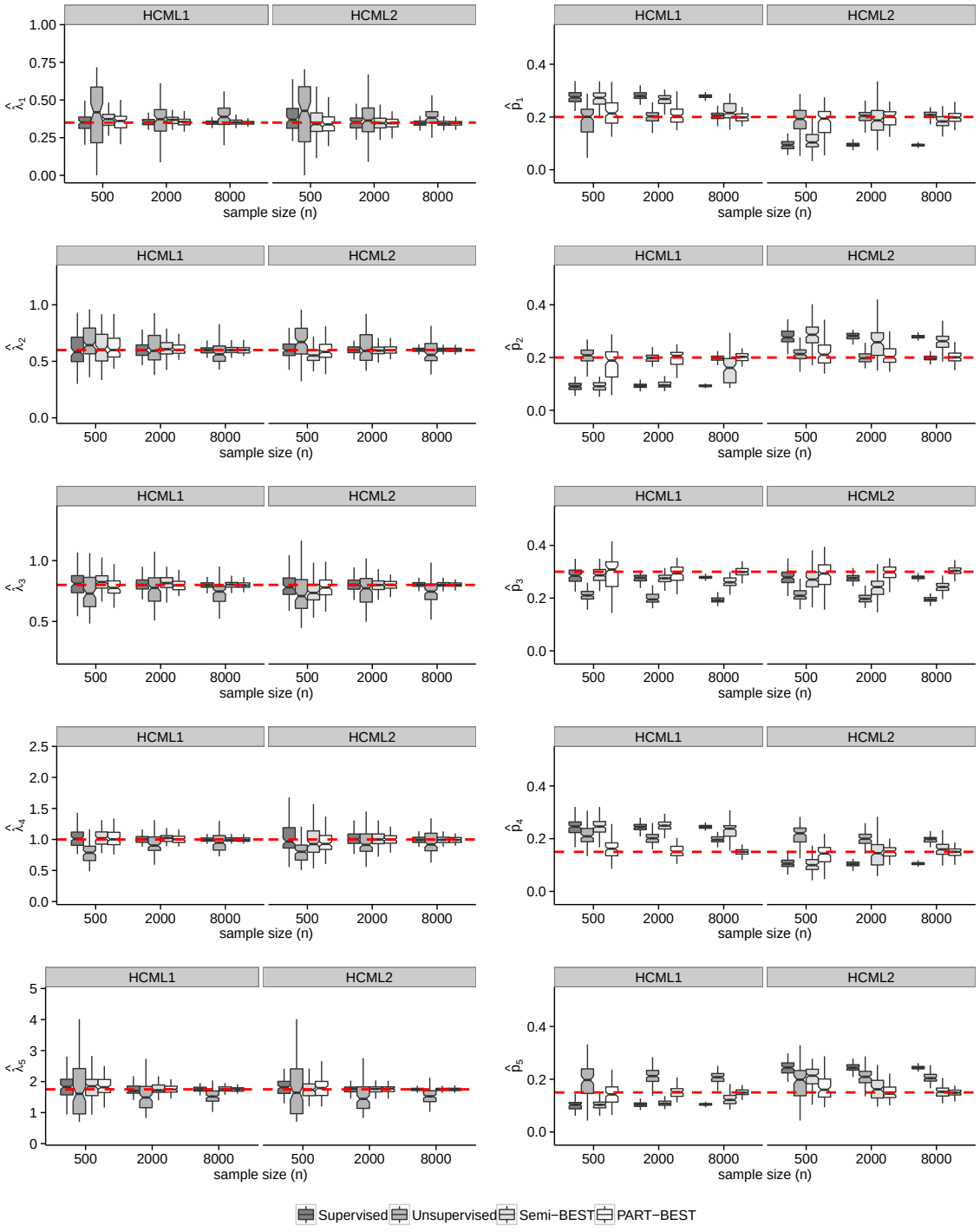


Figure 5.9: (Continued)

Comparison of MLEs on synthetic data sets with U(2,4) censoring time

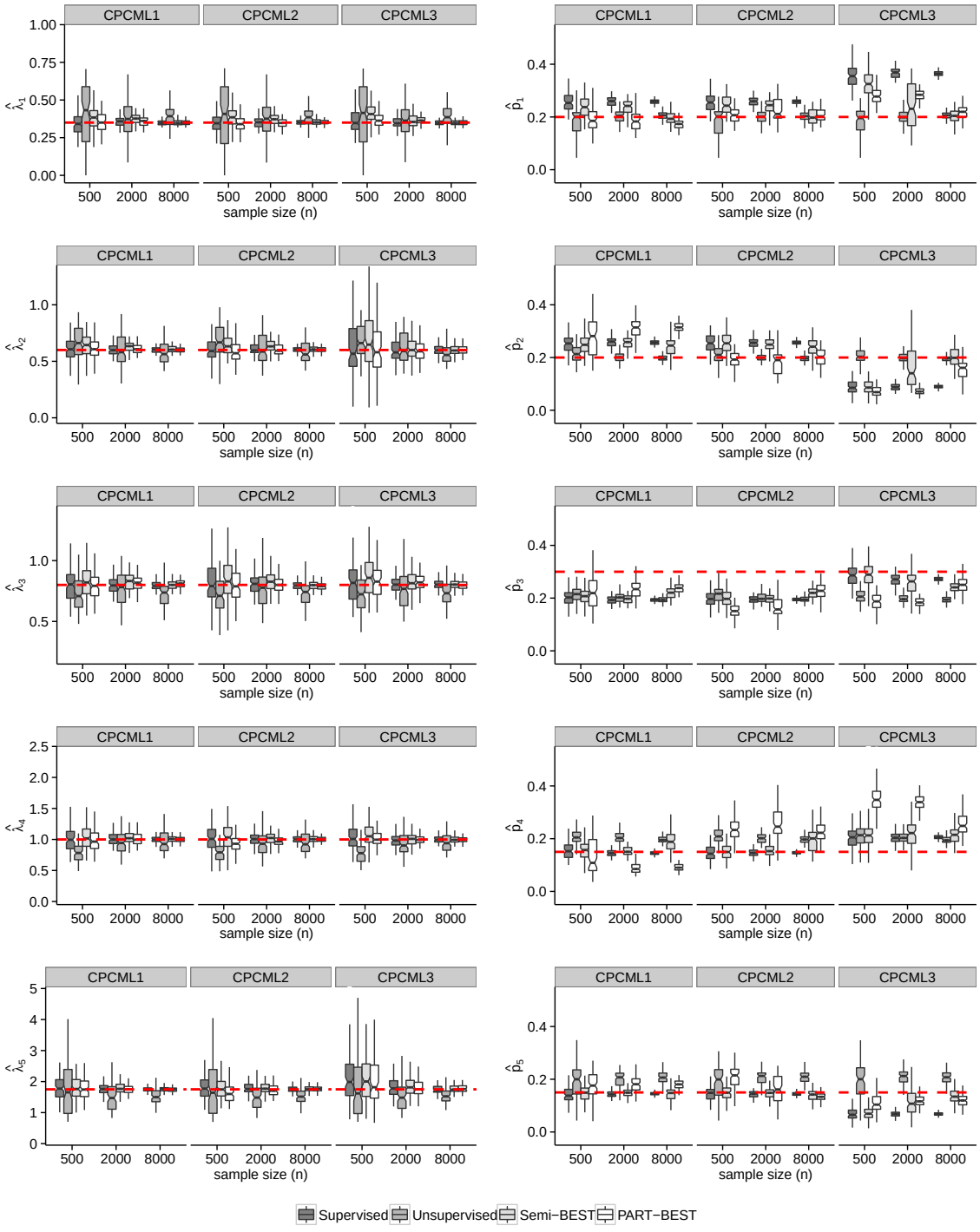


Figure 5.9: (Continued)

Comparison of MLEs on synthetic data sets with U(2,4) censoring time

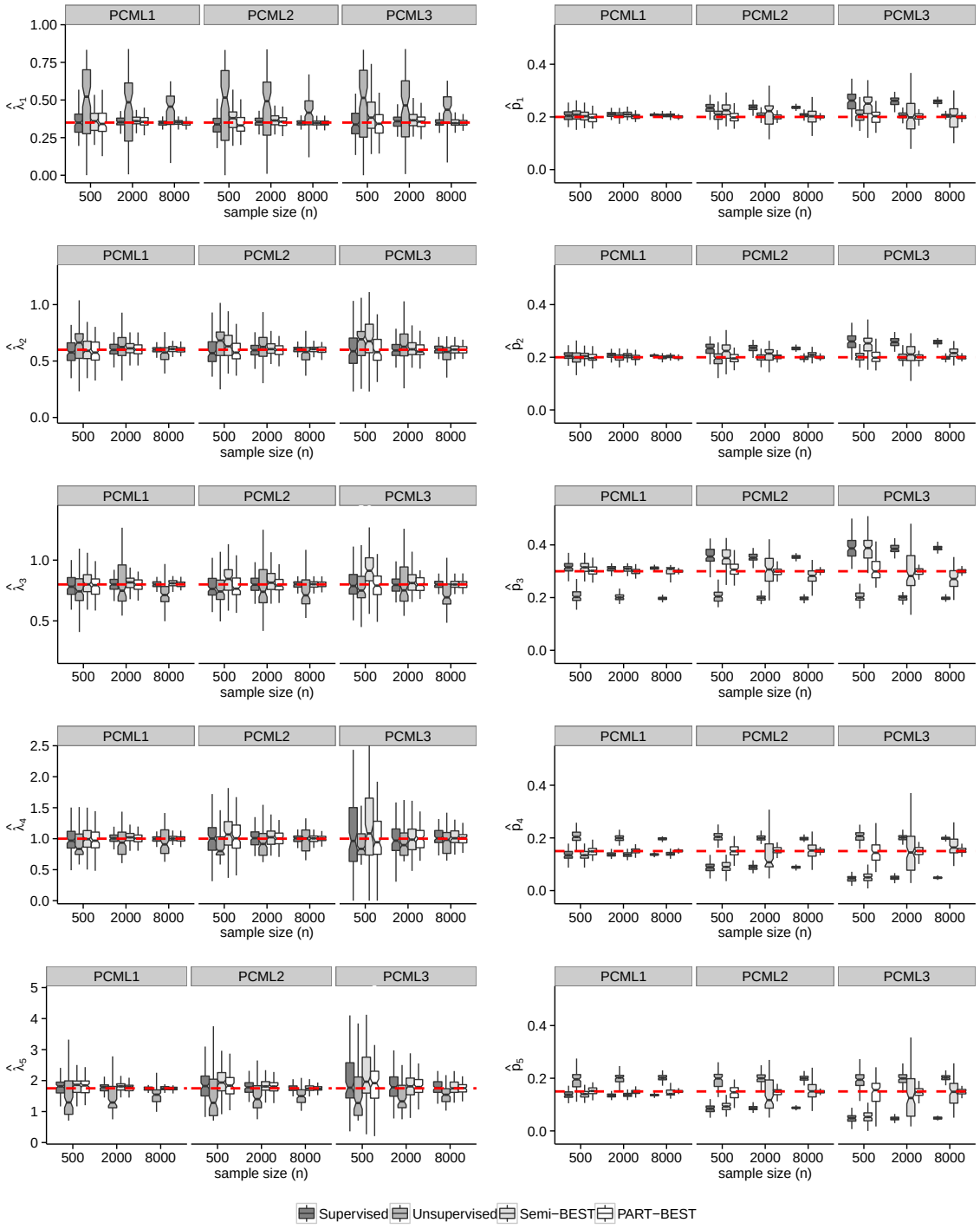


Figure 5.10: Comparison of MLEs on synthetic data sets with U(0.25,1.25) censoring time

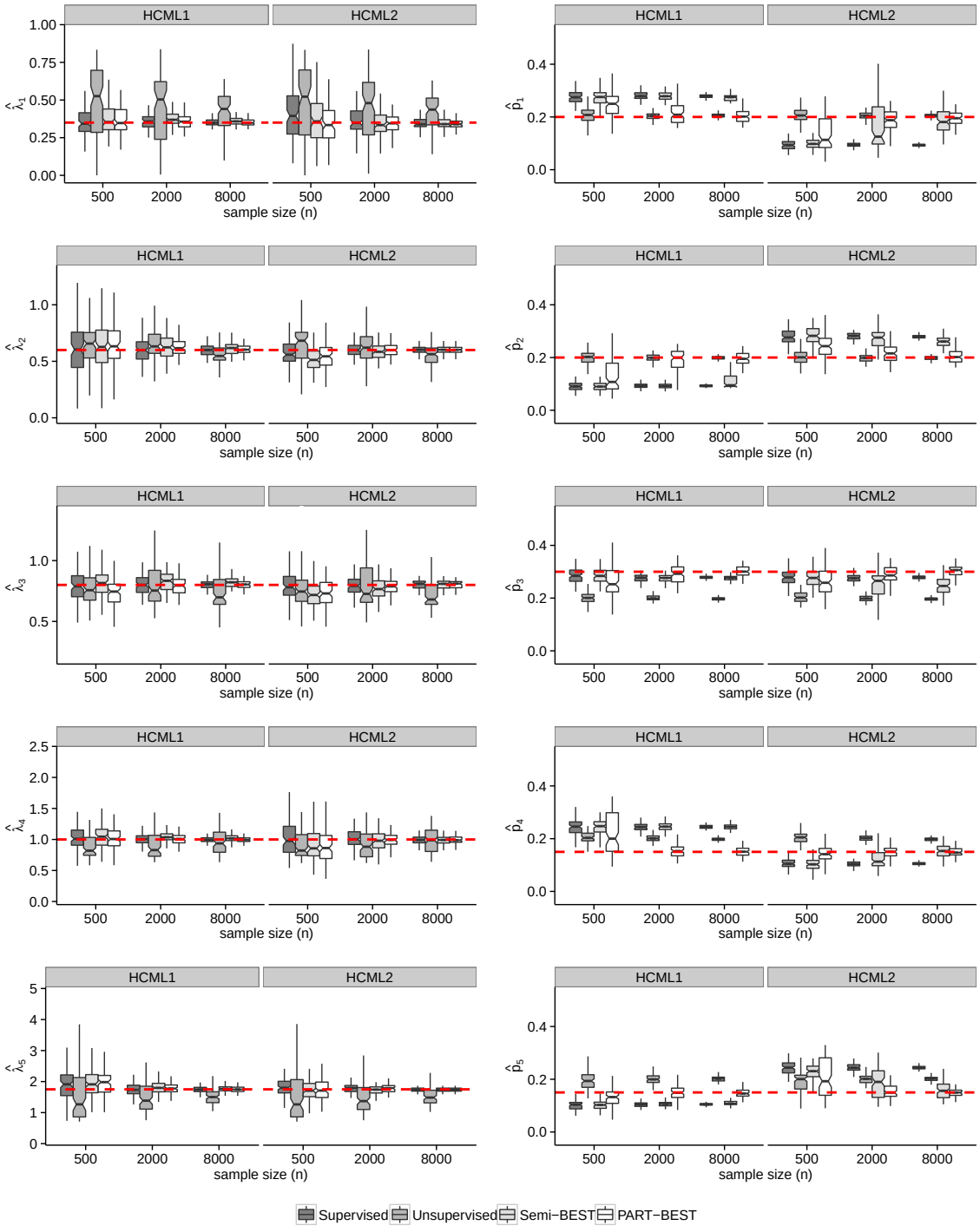


Figure 5.10: (Continued)

Comparison of MLEs on synthetic data sets with U(0.25,1.25) censoring time

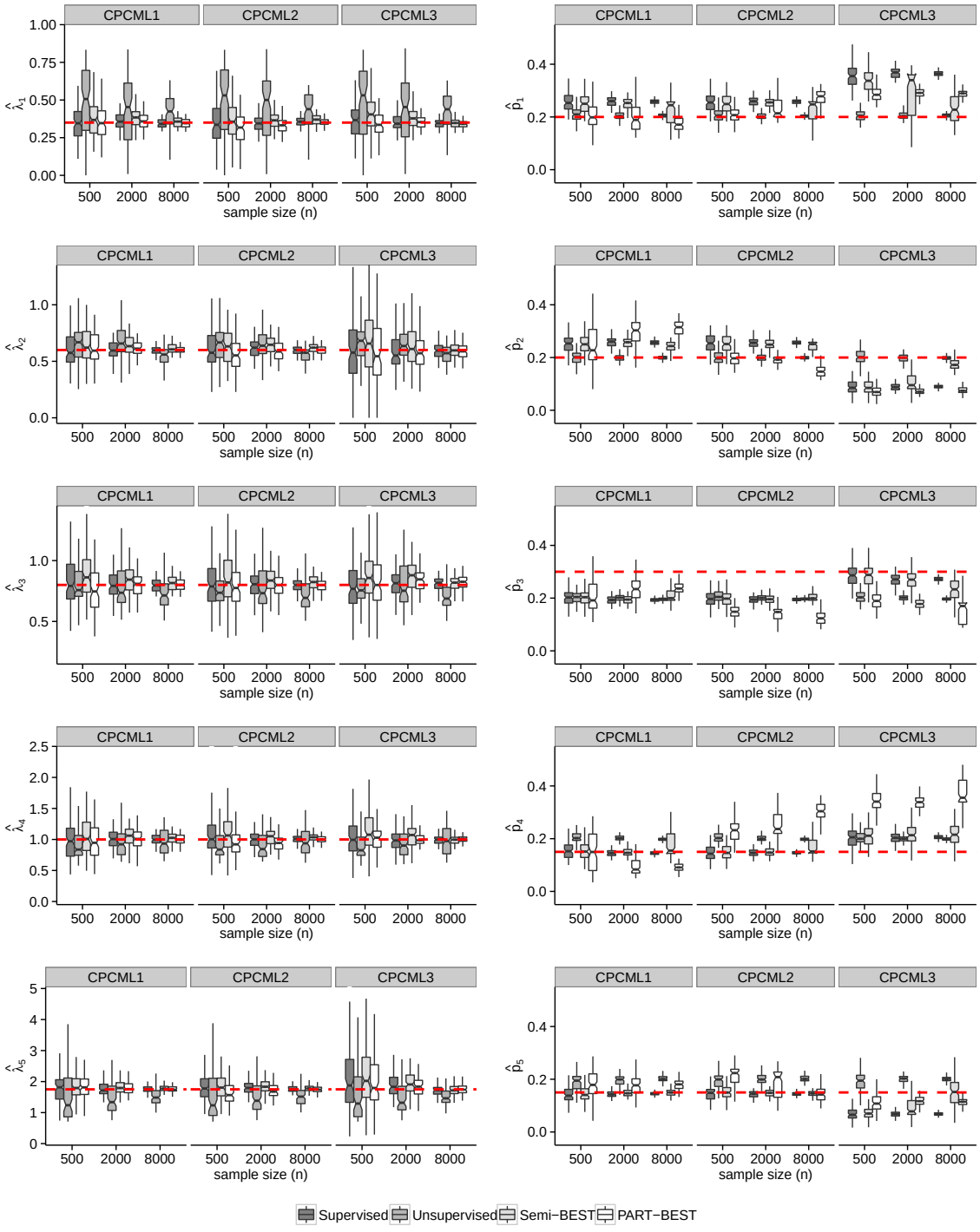


Figure 5.10: (Continued)

Comparison of MLEs on synthetic data sets with U(0.25,1.25) censoring time

supervised learning were consistently smaller than those from unsupervised learning and as small as those from supervised and semi-supervised learning except $\hat{\lambda}_4$ with $n = 500$. The amount of variance reductions in $\hat{\lambda}_k$ were the largest under PCML3 that was designed for the smallest number of precisely labeled data and the largest number of partially labeled data among PMCL mechanisms. MLEs of proportion parameters p_k further emphasized the merits of partially supervised learning. Supervised and unsupervised learning methods produced quite biased estimates of p_k. Semi-supervised learning tended to reduce such bias as a sample size increases. However $\hat{p}_k$ from semi-supervised learning showed very large uncertainties compared to the others. Partially supervised learning consistently produced unbiased estimates whose variances were smaller than or as small as those from the other types of learning methods. For a small sample size ($n = 500$) under PCML3 mechanism partially supervised learning produced $\hat{p}_4$ and $\hat{p}_5$ whose variances were considerably larger than variances of the estimates from the other learning methods. Such demerits have disappeared with larger sample sizes. In this simulation study therefore the proposed partially supervised learning outperformed supervised, unsupervised and semi-supervised learning under PCML mechanisms.

The same interpretations come from the results on data sets under HCML mechanisms. Partially supervised learning consistently obtained less biased MLEs than the other learning methods. A use of partial labels has reduced variances of $\hat{\lambda}_k$ from the results of unsupervised and semi-supervised learning that did not use partial labels. Variances of $\hat{\lambda}_k$ obtained by partially supervised learning were also smaller than those from supervised that only used precisely labeled instances. MLEs of p_k from partially supervised learning showed the largest variance with a small sample size ($n = 500$). For $n = 2000$, partially supervised learning still produced the largest uncertainty in estimating some p_k. For $n = 8000$ sample data sets with U(2,4) censoring time, such uncertainties has been decreased as small as those from semi-supervised learning. With U(0.25,1.25) censoring time however partially supervised learning showed the largest variance of $\hat{p}_k$ even for $n = 8000$. It implies that the proposed partially supervised learning methods possibly produce less robust $\hat{p}_k$ than conventional learning methods with insufficient amount of informative data. Partially supervised learning however consistently found the closest $\hat{p}_k$ to the true value of p_k. We therefore still prefer the results from partially

supervised learning to the others on data sets under HCML mechanisms.

On data sets under CPCML mechanisms, the advantages of using partial labels have been dramatically decreased. For proportion parameters p_k the proposed partially supervised learning algorithms have not produced unbiased estimates even with a large sample size ($n = 8000$). Partially supervised learning often produced significantly more biased estimates than conventional supervised, unsupervised or semi-supervised learning method. In estimating p_k therefore partially supervised learning may not be the best choice of learning methods, unless data contain enough information for partially supervised learning to produce unbiased estimates. On the other side, partially supervised learning still have benefits of precision estimation of component survival distribution parameters λ_k. From the box-plots for $\hat{\lambda}_k$ under CPCML1, CPCML2 and CPCML3 in Figures 5.9 and 5.10, we can observe that PART-BEST consistently produced unbiased estimates with $n = 2000$ and $n = 8000$, while the unsupervised learning method led to biased estimates. Compared to supervised or semi-supervised learning, PART-BEST obtained $\hat{\lambda}_k$ with less uncertainties. We therefore claim that the proposed partially supervised learning methods are still beneficial to obtain MLE of each component survival time distribution even they fail to obtain robust MLE of the mixture model.

A possible reason that partially supervised learning have not outperformed conventional learning methods under CPCML mechanisms can be found in the previous section. Although EM-CPCML exactly includes the CPCML mechanism in its learning procedure, the provided data models required more than 8000 instances to obtain unbiased estimate of FMM by using EM-CPCML. EM-PCML and EM-HCML assumed biased data models from CPCML that lead to biased MLEs. However they improved precision of MLEs from EM-CPCML, because they used less number of parameters. We however can expect that EM-CPCML produces asymptotically unbiased estimates of FMM. A sufficient number of data to obtain unbiased estimates may depend on learning algorithms as well as data models. Comparison of learning methods based on such required number of data would be helpful in evaluating the performances of partially supervised learning algorithms compared to the conventional methods. We leave such issues for further studies.

5.2 A Case Study on Surveillance Data of Gastric Cancer

In this section we conduct a study of partially supervised learning of survival time distributions for gastric cancer patients by using a real clinical database that is provided by Surveillance, Epidemiology and End Results (SEER) program (www.seer.cancer.gov) of the National Cancer Institute. We intend to find how the level of lymph nodes involvement affects survival time of gastric cancer patient.

5.2.1 Data description

Among various kinds of gastric cancer, we select a specific histologic type called *signet ring cell carcinoma*, because it is the most frequently observed specific histologic type of gastric cancer in SEER database. In addition we restrict the first malignant primary site of gastric cancer to be cardia to reduce heterogeneities in a data set except the risk factors in which we are interested. In addition we filtered out cases that were diagnosed before year of 1988. Because the SEER data scheme was dramatically changed in 1988, we had difficulties in integrating all the cases into a single data scheme. We therefore use only cases that have been diagnosed since 1988, which still major proportion of SEER database. For the population that we are interested in, 832 cases have been collected from 1988 to 2008, while only 107 cases have been collected until 1987.

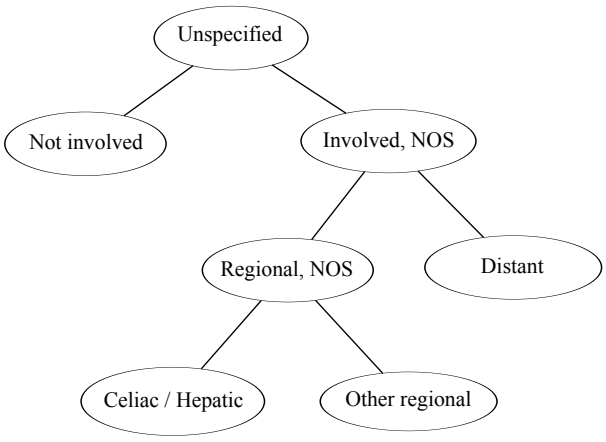


Figure 5.11: Attribute value taxonomy of lymph nodes involvement by gastric cancer tumor

Table 5.4: Number of cases corresponding to each level of the lymph node involvement. (*NOS: not otherwise specified)

Farthest lymph nodes involved	Number of cases	Dead due to cancer	Right-censored
Unspecified	212	182	30
Not involved	192	115	77
Involved, NOS*	4	4	0
Regional, NOS*	144	118	26
Celiac/Hepatic	46	40	6
Other regional	178	137	41
Distant	56	51	5
Total	832	647	185

Data scheme on the lymph nodes involvement attribute was changed in 2004 to separate cases of hepatoduodenal lymph nodes involvement from other celiac / hepatic lymph nodes involvement cases. We however could find only one case explicitly indicating that hepatoduodenal lymph nodes were involved. Such extremely small number of precisely labeled data cannot sufficiently represent a particular class; it can be an obstacle to find reliable estimates of survival time distribution for the class. We therefore use the data scheme that has been used from 1988 to 2003 (Figure 5.11) with migrating data collected since 2004 to the old version of data scheme. One hepatoduodenal case and 16 other celiac / hepatic cases collected since 2004 were labeled `Celiac/Hepatic` in this case study. We indexed classes as follows:

- class 1: Not involved,
- class 2: Celiac/Hepatic,
- class 3: Other regional,
- class 4: Distant.

Table 5.4 describes the 832 selected cases of the gastric signet ring cell carcinoma on the cardia. Actual survival time of 647 cases has been observed, while 185 cases were right-censored either because a patient is still alive or because a patient died of other causes than cancer. In addition while 472 cases were precisely labeled based on AVT in Figure 5.11, 360 cases were imprecisely labeled. So we can expect that improving precisions of labeling for those 360 cases

will lead to better estimates of the survival time distribution within each of four classes of lymph nodes involvement. In particular 148 cases of `Involved,NOS` and `Regional,NOS` may be fully utilized only by partially supervised learning methods. We expect fully utilizing 148 partially labeled data leads to more reliable knowledge about the risks of lymph nodes involvement.

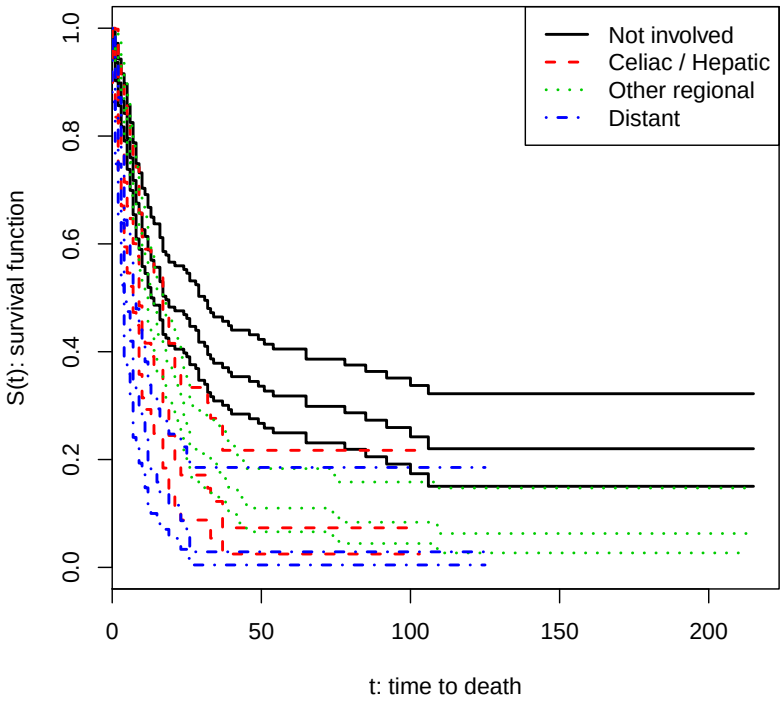


Figure 5.12: Kaplan-Meier curves for patients with the gastric signet ring cell carcinoma on the cardia, with 95% confidence intervals.

Figure 5.12 shows Kaplan-Meier curves (Kaplan and Meier, 1958) corresponding to the subpopulations, which were estimated by using only 472 cases whose origins were exactly known based on the hierarchical structure of observable labels in Figure 5.11. The inner curves represent the expected probabilities of survival longer than time t within precisely labeled `Not involved` cases (black solid line), `Celiac/Hepatic` cases (red dashed line), `Other regional` case (green dotted line), and `Distant` cases (blue dot-dash line). The outer two curves represent 95% confidence intervals of the survival time distribution at each risk. The difference between

expected survival rates shows that the lymph node involvement is an important risk factor for a death caused by the gastric cancer. However, large uncertainties of the estimation can be barriers to clarify the effects of lymph nodes involvement on survival time. This study will make the incomplete diagnostic information in the remaining 360 cases contribute in improving precision of the expected survival time at each risk level.

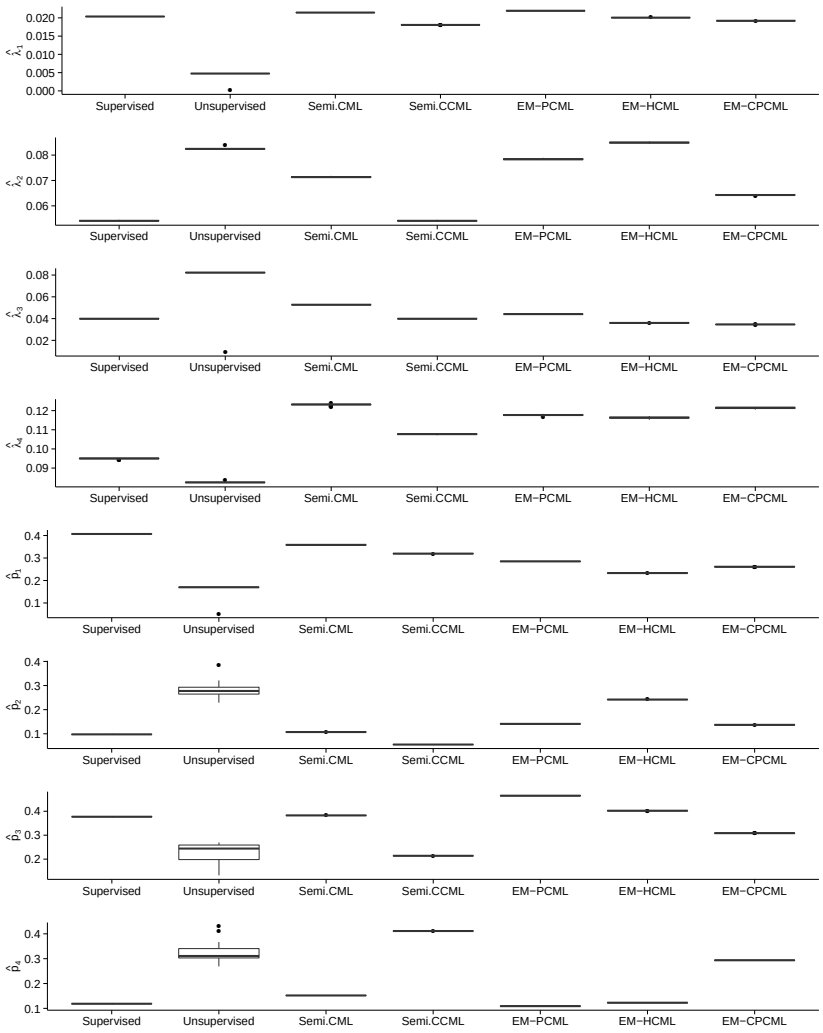


Figure 5.13: Box-plots of MLEs on 20 random noise sets

In SEER database survival time is an integer value with a measurement unit of month. Therefore the data set is inappropriate to estimate exponential survival time distributions that assume continuous time rather than discrete one. Moreover SEER database set survival time

to be zero month for very short time to death, which makes MLE cannot be found. Although there are existing approaches to estimate survival time distributions with such observed survival time intervals, we do not follow such methods because the issue of imprecisely observed survival time is not of our primary interest. We rather simply overcome the feasibility problems by adding small positive random noises to observed survival time; we added a random noise generated from a uniform distribution U(0,1) to each observed or censored survival time in SEER database. Although our approach leads to slightly biased estimates from the estimates learned from interval values, we supposed such bias is ignorable compared to differences in estimates between learning algorithms. We have generated 20 data sets with independent random noises in survival time to show robustness of the results for random noises. Figure 5.13 represents MLEs of FMM obtained from each learning method on 20 data sets with randomly noised survival time. Very small variation of estimates within algorithm compared to differences between algorithms implies that random noise effects on survival time data will not affect the preference of learning algorithms on the given data. We therefore add small random noises on survival time data without expected critical effects on the results of estimations based on interval survival time data.

5.2.2 Exploratory data analysis

In contrast to simulations on synthetic data in Section 5.1 we do know what the true finite mixture distribution is. It is therefore impossible to evaluate MLEs based on the agreement with the true parameter values. In addition the underlying missing label mechanism is not known, so we cannot evaluate whether AIC-based model selection agrees with the true knowledge. However we may find some hints about the true model by conducting exploratory analysis of SEER data with the some attributes that have not been used in the FMM learning. Because only four cases that were labeled `Involved,NOS` do not significantly affect the results, we have focused on exploratory data analysis for 212 `Unspecified` cases and 144 `Regional,NOS` cases.

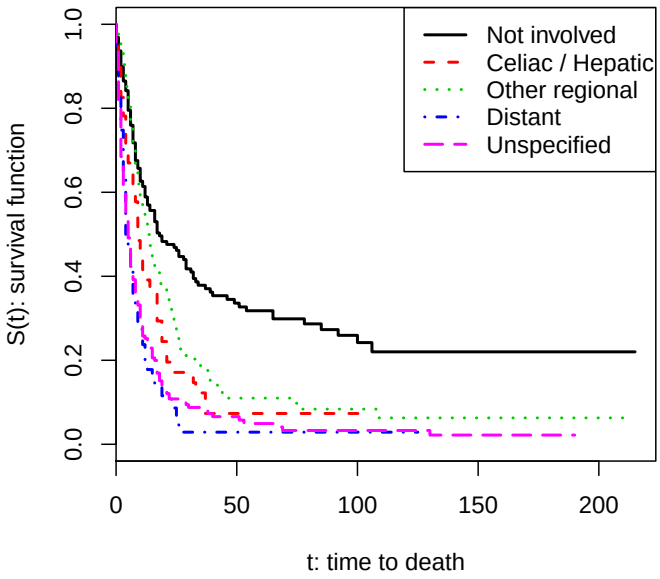


Figure 5.14: Comparison of Kaplan-Meier curve for `Unspecified` cases to Kaplan-Meier curves for precisely labeled cases

5.2.2.1 Missing label mechanisms on unspecified data

212 cases that were labeled `Unspecified` are supposed to belong to any of four classes. When we compare nonparametric survival distribution within `Unspecified` cases to those for precisely labeled cases by using Kaplan-Meier curves, the survival patterns within `Unspecified` cases was very close to the survival patterns of `Distant` cases (Figure 5.14). In particular, survival patterns up to 20 months which cover over than 80% of `Unspecified` cases were almost the same to each other. We therefore can expect that `Distant` class takes a major proportion of `Unspecified` cases, while only 12% of precisely labeled cases have been labeled as `Distant`.

To assure that a major proportion of `Unspecified` cases belongs to `Distant` class, we have investigated how the lymph nodes involvement is correlated with the following two attributes that have not been used in defining classes:

Table 5.5: SEER coding system for EOD extension (1988-2003)

Code	Description
00, 05	Noninvasive
10–16	Invasive tumor confined to a specific inner stomach layers
20	Muscularis propria invaded
30	Localized, NOS
40	Extension to/through wall
45	Extension to adjacent (connective) tissue
50	Invasion of/through serosa
55	(45) + (50)
60	Extended to adjacent structures (e.g. liver, diaphragm, esophagus, duodenum)
70	Extended to abdominal wall, retroperitoneum, kidney, and/or adrenal gland
80	FURTHER contiguous extension
85	Metastasis
99	UNKNOWN if extension or metastasis

Table 5.6: SEER coding system for historic stage A

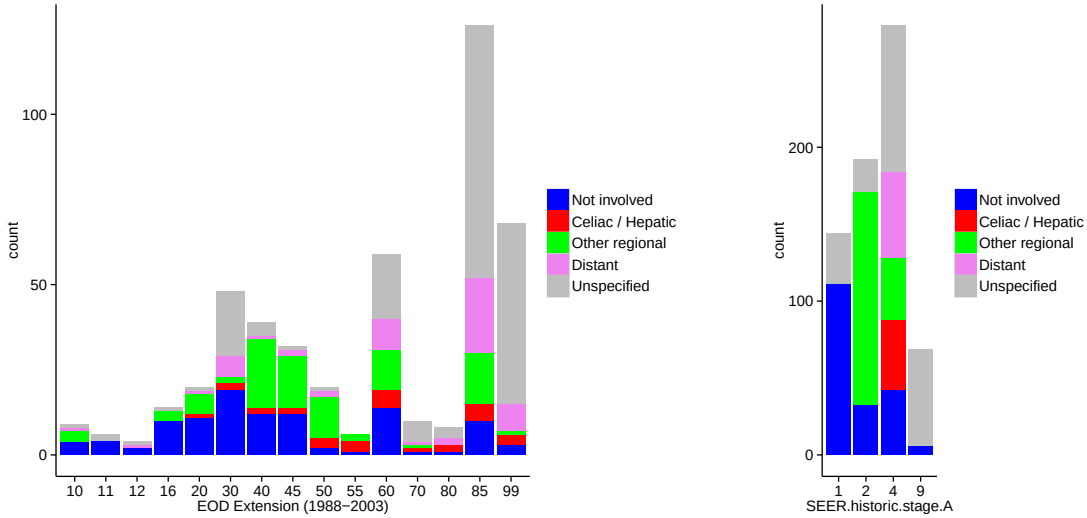
Code	Description
0	In situ - noninvasive
1	Localized - confined entirely to the organ of origin
2	Regional- extended into surrounding organs or tissues and/or regional lymph nodes
4	Distant - extension or metastasis to distant organ / includes distant lymph nodes
9	Unstaged

- **EOD extension**: the farthest documented extension of tumor away from the primary site, either by contiguous extension or distant metastases,
- **SEER historic stage A**: a stage of cancer that collapses the detailed EOD information collected by SEER.

The `EOD extension` represents important diagnostic information that the lymph nodes involvement does not completely explain. However the EOD extension and the lymph nodes involvement are expected to be strongly correlated, because both of them represent how far the cancer tumor has been extended. The `SEER historic stage A` represents a severity of cancer based not only on lymph nodes involvement but also on extensions of tumor and metastasis as well. A data filed of the EOD extension in SEER database has been valid until 2003 but has not been used since 2004. We therefore use only 596 cases that have been collected until 2003 for exploratory analysis of correlation between the EOD extension and the lymph nodes involvement. On the contrary the SEER historic stage A has been coded for all the 832 cases that have been used in this study. The SEER coding system of the EOD extension and the SEER historic stage A are described in Tables 5.5 and 5.6, respectively. A common characteristic of two coding systems is that the severity of cancer increases as the coded number increases. Code 99 for the EOD extension 9 for the SEER historic stage A represent unknown cases on each attribute. More details of the SEER coding system can be found on the SEER website www.seer.cancer.gov.

Figure 5.15(a) shows the distribution of EOD extension values within `Unspecified` group is very similar to the distribution for `Distant` group. In particular, code 85 on the EOD extension is most frequently observed within each of `Distant` group and `Unspecified` group, while other values are more frequently observed within the other classes: 30 within `Not involved`, 60 within `Celia / Hepatic`, and 40 within `Other regional`. So the largest proportion of `Unspecified` cases is expected to belong to the `Distant` class.

Because the staging relied on lymph nodes involvement, we can find a strong correlation between lymph nodes involvement and SEER historic stage in Figure 5.15(b). In particular we observe that all the cases of `Celiac/Hepatic` and `Distant` groups are matched to code 4



(a) EOD extension (b) SEER historic stage A

Figure 5.15: Distributions of (a)EOD extension and (b)SEER historic stage A within `Unspecified` cases as well as each group of precisely labeled cases

of SEER historic stage. We also can see that the major proportion of `Unspecified` lymph node involvement cases has been staged 4. Moreover in Figure 5.16 60% of the earliest died of unstaged cases (stage 9) show a death rate as high as cases of stage 4. We therefore expect that a major proportion of stage 9 would be belonging to stage 4. From those observations, we can expect that a major proportion of `Unspecified` cases involve `Celiac/Hepatic` or `Distant` classes. On the other hand a considerable proportion of `Unspecified` group is corresponding to stage 1 that only appears for `Not involved` group. We therefore expect that a considerable number of `Unspecified` cases do not involve lymph nodes in cancer tumor.

The above observations consistently lead to an implication that "`Distant` class covers the largest proportion of 212 `Unspecified` cases," which violates the assumption of common missing label mechanism in the Semi-CML algorithm, the assumption of pattern-conditional missing label mechanism in the EM-PCML algorithm, and the assumption of hierarchy-conditional missing label mechanism in the EM-HCML algorithm on the given data set.

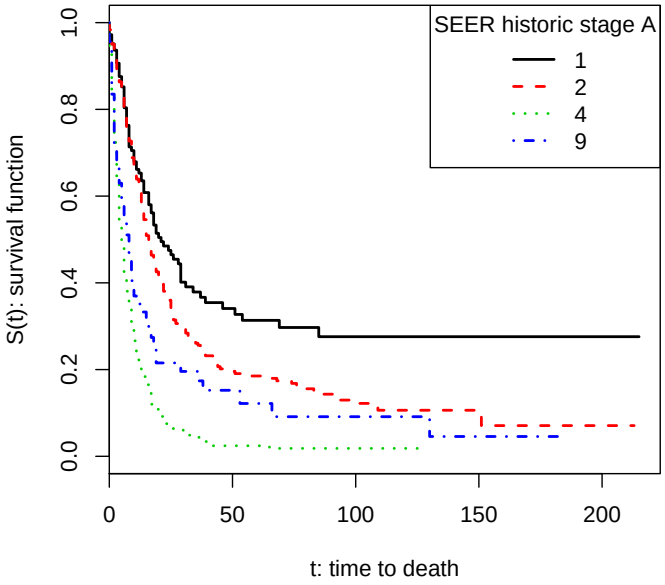


Figure 5.16: Kaplan-Meier curves depending on SEER historic stage A for gastric signet ring cell carcinoma on the cardia

5.2.2.2 Missing label mechanisms on the `Regional,NOS` group

Now we conduct the same exploratory data analysis for 144 cases of the `Regional,NOS` group. For a case that was labeled by `Regional,NOS` the most specific information about lymph node involvement is supposed to be either `Celiac/Hepatic` or `Other regional`. Figure 5.17 represents Kaplan-Meier (KM) curve for the `Regional,NOS` group and KM curves for precisely labeled two groups that are possible true classes for `Regional,NOS` cases. To show how the difference in survival rates between `Celiac/Hepatic` class and `Other regional` class

Table 5.7: p-values of statistical tests for differences in survival patterns between `Celiac/Hepatic` and `Other regional` groups

Test	p-value
Mental-Haenszel	0.0913
Gehan-Wilcoxon	0.0373

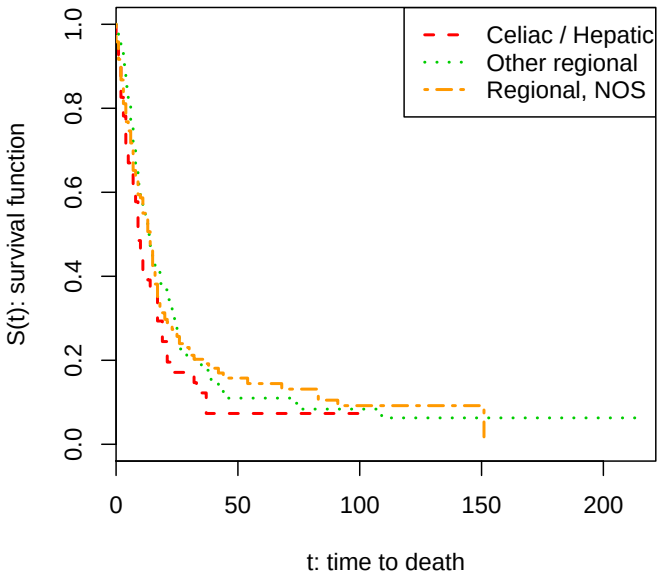
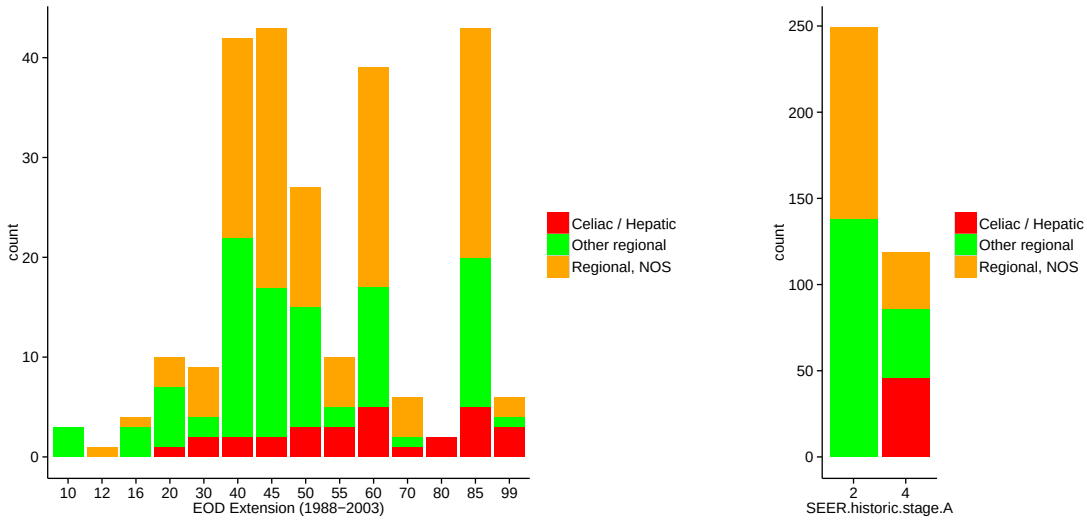


Figure 5.17: Comparison of Kaplan-Meier curve for `Regional,NOS` cases to Kaplan-Meier curves for precisely labeled data that regional lymph nodes were involved

is reliable, we conducted Mental-Haenszel log-rank test and Gehan-Wilcoxon test by using `survdiff` function that has been implemented in `survival` package of `R`. It is noteworthy that Gehan-Wilcoxon test is good to detect early differences between two groups, while Mental-Haenszel test is used for testing differences in overall survival rates. The p-values from the statistical tests (Table 5.7) imply that the two classes have significantly different survival rates in early periods, while overall survival rates across the whole time periods are less significant. In Figure 5.17 the survival rate within `Regional,NOS` group is closer to the survival rate of `Other regional` class during the first 30 months and lies at the middle of two precisely labeled groups after the the 30th month. By combining observations from Table 5.7 and Figure 5.17 we can have an insight that about 80% of `Regional,NOS` cases who were or might be died in first 30 months are more likely to be belonging to the `Other regional` class rather than the `Celiac/Hepatic` class.

It was unclear whether the distribution of EOD extension within `Regional,NOS` group is



(a) EOD Extension (b) SEER historic stage A

Figure 5.18: Distributions of (a)EOD extension and (b)SEER historic stage A within `Regional,NOS` cases as well as precisely labeled regional lymph nodes involvement cases

closer to that of `Celiac/Hepatic` class or that of `Other regional` in Figure 5.18(a). We however observed that SEER historic stage for 77% `Regional,NOS` cases was stage 2 that has not come with `Celiac/Hepatic` group. It is therefore expected that the major proportion of the `Regional,NOS` group is belonging to the `Other regional` class.

The observations of the KM curves and the SEER historic stages imply that "`Other regional` class covers the largest proportion of 144 `Regional,NOS` cases," which still makes assumptions of the following missing label mechanisms plausible: common missing label mechanism, pattern-conditional missing label mechanism, or hierarchy-conditional missing label mechanism.

5.2.3 Estimated survival time model

From exploratory data analysis in Section 5.2.2, we have built the following two reasonable implications:

- **Implication 1** - `Distant` class covers the largest proportion of 212 `Unspecified` cases;

- **Implication 2** - `Other regional` class covers the largest proportion of 144 `Regional,NOS` cases.

In this section we evaluate which learning algorithm most agrees with such implications.

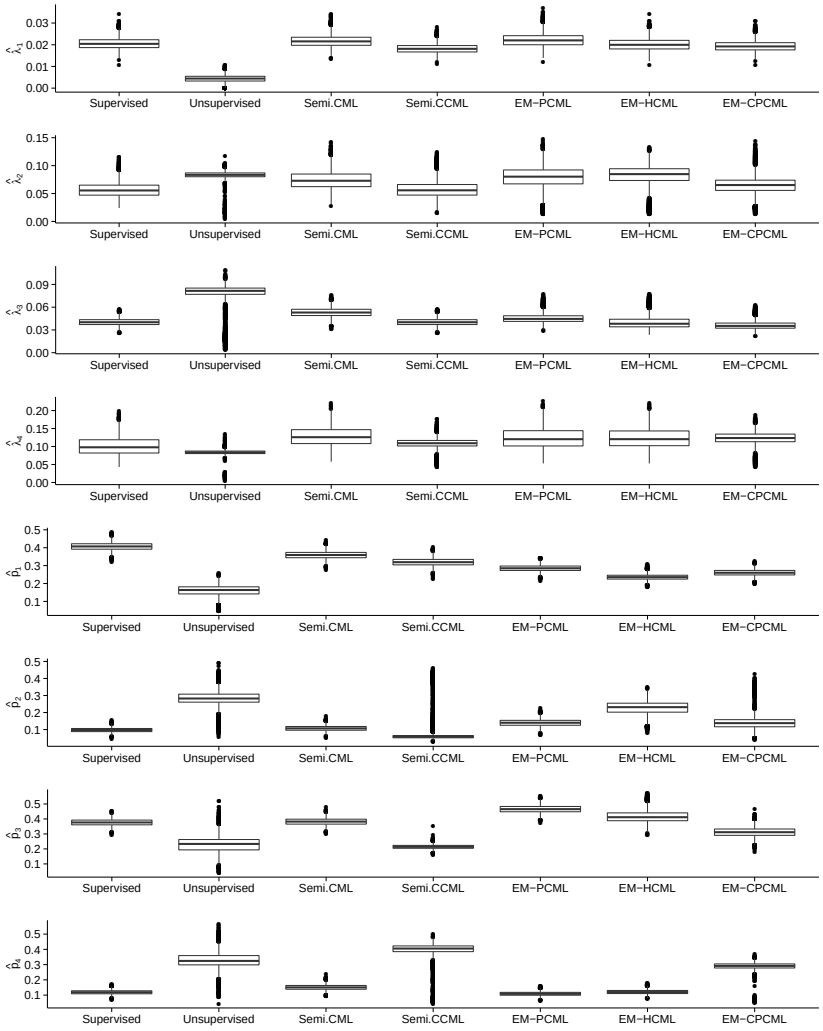


Figure 5.19: Box-plots of MLEs on 4000 bootstraps

To avoid sampling bias, we have generated 200 bootstraps from each of 20 randomly noised data sets. As a result we had estimated finite mixture models on each of 4000 bootstraps (Figure 5.19). Distributions of $\hat{\lambda}_2$, $\hat{\lambda}_3$ and $\hat{\lambda}_4$ from unsupervised learning were almost identical; it does not agree with our observations on Kaplan-Meier curves in Figure 5.12 that shows significant differences of the survival rate for `Distant` group from the others. Thus we ignore

Table 5.8: Averaged expectations of class proportions within the `Unspecified` group for 20 randomly noised data sets

	Class			
Method	Not involved	Celiac/Hepatic	Other regional	Distant
Semi-CML	0.26	0.13	0.39	0.22
Semi-CCML	0.16	0.00	0.00	0.84
EM-PCML	0.21	0.18	0.45	0.16
EM-HCML	0.01	0.35	0.44	0.20
EM-CPCML	0.12	0.00	0.00	0.88

Table 5.9: Averaged expectations of class proportions within the `Regional,NOS` group for 20 randomly noised data sets

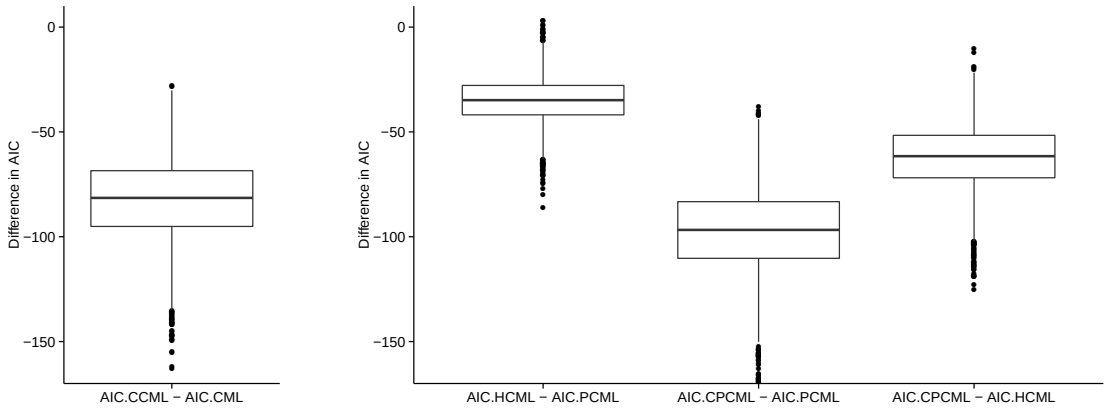
	Class			
Method	Not involved	Celiac/Hepatic	Other regional	Distant
Semi-CML	0.35	0.11	0.38	0.16
Semi-CCML	0.27	0.00	0.00	0.73
EM-PCML	-	0.23	0.77	-
EM-HCML	-	0.56	0.44	-
EM-CPCML	-	0.45	0.55	-

the results from unsupervised learning hereafter.

Our first implication is that $\hat{p}_4$ should be much higher than 12%, which is the proportion of `Distant` group among 472 precisely labeled cases. Figure 5.19 shows that $\hat{p}_4$ from supervised learning, Semi-CML, EM-PCML and EM-HCML is still very close to the proportion within precisely labeled cases, while distributions $\hat{p}_4$ from Semi-CCML and EM-CPCML are centered around 30%. Table 5.8 more directly guides the agreement of the estimates with the first implication by representing how each learning method expects the class proportion within cases that whose true classes were unspecified. Among two semi-supervised learning algorithms and three partially supervised learning algorithms, only Semi-CCML and EM-CPCML has agreed to the first implication.

Table 5.9 shows the agreement of each learning methods with our second implication on the given data set. To agree with the second implication the expected proportion of `Other regional` should be significantly higher than the expected proportion of `Celiac/Hepatic`. The results from Semi-CML and EM-PCML strongly agreed with the second implication, while EM-

CPCML moderately agreed. A fundamental difference between semi-supervised learning and partially supervised learning is that semi-supervised learning algorithms still expected some proportions of `Regional,NOS` group belongs to `Not involved` or `Distant` classes, which is not a valid expectation. In particular, Semi-CCML did give extremely small expectations to the proportions of `Celiac/Hepatic` and `Other regional` classes whose sum is supposed to be one in valid expectations. We therefore conclude that partially supervised learning outperforms semi-supervised learning from the perspective of our second implication.



(a) Semi-supervised learning (b) Partially supervised learning

Figure 5.20: AIC-based comparison of the results within (a)semi-supervised learning methods and (b)partially supervised learning methods for 4000 bootstraps

Figure 5.20 represents AIC-based comparisons of the results from different learning algorithms within (a)semi-supervised learning and (b)partially supervised learning. For all the 4000 bootstraps Semi-CCML consistently outperformed Semi-CML. EM-CPCML also outperformed other partially supervised learning algorithms on all the 4000 bootstraps. Even Semi-CML and EM-PCML showed better agreement with our second implication than Semi-CCML and EM-CPCML respectively, their poor agreements to our first implication demerited them to be selected as the best model within each type of learning methods. Although the results from both Semi-CCML and EM-CPCML strongly agreed with our first implication, they have shown extremely different agreements to the second implication. EM-CPCML expected 55%

Table 5.10: Estimates of a finite mixture model obtained by EM-CPCML on SEER data: mean, median, 2.5th percentile and 97.5th percentile that were simulated from 4000 bootstraps

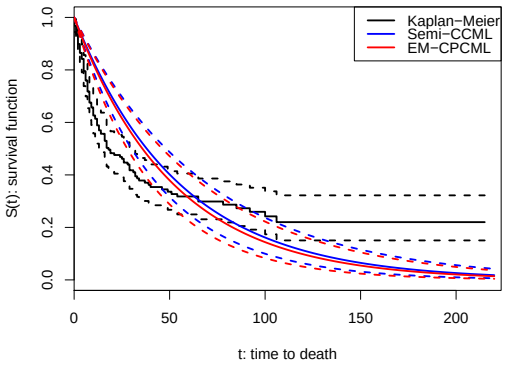
Parameter	Mean	Median	2.5th percentile	97.5th percentile
λ_1	0.019	0.019	0.015	0.025
λ_2	0.064	0.065	0.024	0.106
λ_3	0.037	0.035	0.028	0.053
λ_4	0.123	0.124	0.070	0.155
p_1	0.260	0.260	0.225	0.296
p_2	0.143	0.138	0.066	0.324
p_3	0.312	0.312	0.247	0.383
p_4	0.284	0.292	0.076	0.330

of partially labeled `Regional,NOS` cases would belong to `Other regional`; its agreement to the second implication was not strong but still agreed rather than disagreed. In addition EM-CPCML expected the remainder would belong to `Celiac/Hepatic`, which is a valid expectation for partial label `Regional,NOS`. On the other side Semi-CCML provided invalid expectations and disagreed with our second implication. We therefore conclude that EM-CPCML is the best algorithm for learning a mixture of exponential survival time distribution on the given SEER data set.

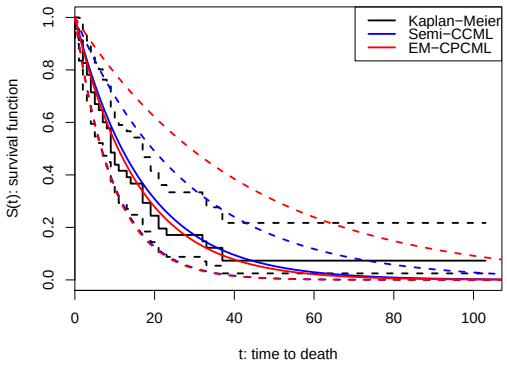
Estimates of FMM obtained by EM-CPCML on 4000 bootstraps are summarized in Table 5.10. In addition we compared the estimated survival functions to those from Semi-CCML as well as Kaplan-Meier curves on precisely labeled data in Figure 5.21. We used the 97.5th percentile of $\hat{\lambda}_k$ for the lower bound survival function, mean for center, and the 2.5th percentile for upper bound. We summarize several findings and implications from the comparison.

- Survival time for `Not involved` class did not look exponentially distributed. However Kaplan-Meier curves in the other classes fairly matched the estimated exponential distributions. Modeling survival function for class 1 with different types of distribution may improve the reliability of the estimated mixture models.

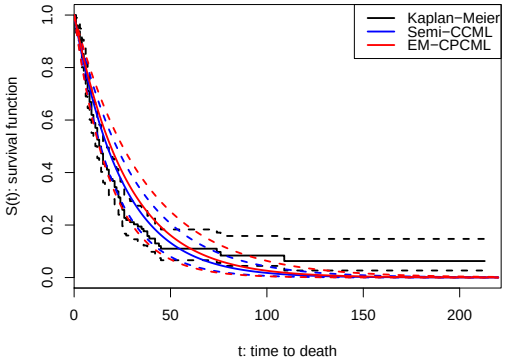
- Semi-CCML estimated FMM with less variance than EM-CPCML. It means benefits of reducing variances by using additional data were not enough to countervail increased variances caused by increasing complexity of the model.



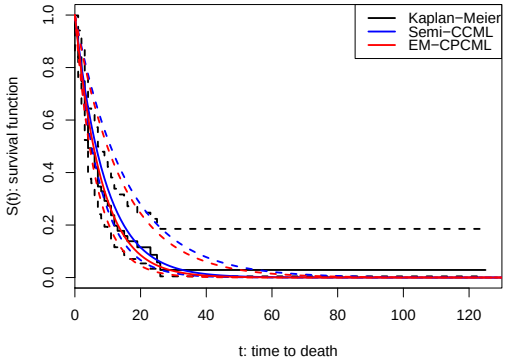
(a) Not involved



(b) Celiac/Hepatic



(c) Other regional



(d) Distant

Figure 5.21: Comparison of estimated survival functions to Kaplan-Meier curves

- EM-CPCML slightly better agreed with Kaplan-Meier curves than Semi-CCML did. It might be because EM-CPCML outperformed Semi-CCML in computing expected memberships of `Regional,NOS` cases to each class. Semi-CCML however better agreed with Kaplan-Meier curve for `Other regional` class. A plausible reason for this phenomenon is that EM-CPCML estimated that too much proportion of the `Regional,NOS` group belongs to `Celiac/Hepatic` class. Assigning `Regional,NOS` cases with short survival time to `Celiac/Hepatic` class caused the overestimation of survival chances in `Other regional` class, which caused bias.

5.3 Summary

In this chapter we conducted a numerical simulation study as well as a case study on SEER data. Both studies clearly showed benefits of using the proposed partially supervised learning methods compared to conventional supervised, unsupervised, and semi-supervised learning. In our simulation study partially supervised learning methods outperformed or performed as well as the others unless information from partially labeled data is too small to countervail increased model complexity. CPCML mechanisms often required more samples than PCML or HCML mechanisms to obtain reliable estimates from partially supervised learning. In a case study we EM-CPCML best agreed with our implications from exploratory data analysis. Although we could not reduce uncertainties of estimates from semi-supervised learning, we still could reduce bias of estimates by using additional valid data that semi-supervised learning has not used. It is supposed that there was model bias caused by a wrong assumption on survival time distribution for `Not involved` class. Reducing such bias is desired to fairly evaluate the performance of partially supervised learning to the conventional learning on the given data set.

In next chapter we conclude our contributions from findings in this study and discuss potentially valuable future works.

CHAPTER 6. CONCLUSIONS

Medical surveillance data is composed of observed survival time of patients and covariates that characterize patients. We are usually faced with a lot of data that have only partially specified values on such covariates when we want to learn how patients' characteristics affect their survival time from a surveillance database. Partially specified class values may be caused by either incomplete data-intake practices or changes in data-intake procedures. Despite of potential advantages of using such partial labels like reduced bias or uncertainties in estimating the relations between class covariates and survival time, a systematic study of learning from partial labels has not been popularly conducted in the field of medical data mining. The main contribution of our study is fully utilizing the partially labeled data for learning finite mixture models of survival time distributions.

Throughout this study we have focused on the mechanisms that caused partially specified class labels. By defining four different missing label mechanisms we proposed the following four Expectation-Maximization algorithms for maximum likelihood estimation of finite mixture models:

- EM-OCML: EM algorithm under overall common missing label mechanism,
- EM-PCML: EM algorithm under pattern-conditional missing label mechanism,
- EM-HCML: EM algorithm under hierarchy-conditional missing label mechanism,
- EM-OCML: EM algorithm under class-pattern-conditional missing label mechanism.

In fact EM-OCML has already been described by Ambroise and Govaert (2000). We however made additional contribution by uncovering an underlying assumption on the algorithm in Ambroise and Govaert (2000) and showed feasibility problems that makes EM-OCML ap-

plicable under very restricted conditions. The other three methods (EM-PCML, EM-HCML, EM-OCML) are our original contributions that allowed generalized conditions that EM-OCML cannot deal with. In particular we focused on partial labels that are hierarchically defined by attribute value taxonomy that represents a coding system of a covariate in medical surveillance databases. EM-HCML is only feasible to hierarchically defined partial labels. We emphasize that all the three EM algorithms originally proposed in this study are always feasible to hierarchically defined partial labels, whereas EM-OCML often fails to obtain a feasible solution.

Even though the proposed algorithms are applicable to finite mixtures of any parametric distributions that allow maximum likelihood estimation, we only conducted experiments on finite mixtures of survival time distributions from which this study has been motivated. Specifically we used exponential distributions to represent the survival time distribution within each class. First a simulation study for a mixture of five exponential distributions has been conducted by varying sample sizes and missing label mechanisms. The proposed EM algorithms often outperformed conventional supervised, unsupervised, and semi-supervised methods by reducing bias as well as variance of estimates; it is the advantage what we expected to take by using additional valid information compared to the conventional methods. The proposed algorithms performed well especially when a sufficient number of samples are given. With a limited number of samples however there was still a chance that the conventional methods outperform the proposed methods.

We also conducted a case study for gastric cancer cases collected by Surveillance, Epidemiology and End Results (SEER) program. By considering the farthest area that lymph nodes involved by tumor as a class label, we concluded that EM-CPCML is the best algorithm for learning the finite mixture survival time model on the sampled SEER data. EM-CPCML agreed with our two implications from explanatory data analysis that 1) a case whose status of lymph nodes involvement has been unknown is most likely to involve distant lymph nodes by tumor and 2) a case whose has been known only that regional lymph nodes were involved is most likely to involve regional lymph nodes other than celiac and hepatic. Supervised, unsupervised and semi-supervised methods have been failed to agree with the implications. In the above experiments we found AIC criteria perform well for selecting the best partially supervised algorithm

among our proposed methods.

We conclude with suggesting three important future works for making this research richer.

- **Extension to a variety of survival time models** – In this study we restricted a component survival time model to a univariate survival time distribution with right-censored time. In our case study however we found that attributes that have not been used to produce class labels delivers additional information about survival time that the class labels cannot explain. We therefore expect that a study for learning mixtures of survival time regression models will lead to more reliable survival time models. Also a study for partially supervised learning of non-parametric or semi-parametric mixtures whose component distributions are Kaplan-Meier curves or Cox-regression models, respectively, will be a greatly valuable extension of this study to allow more flexibility in data models.

- **Constraints on parameters from experts** – In our simulation study we added an ordering constraint to unsupervised learning method, which incorporated the true knowledge in estimation. The results from the proposed algorithms disagreed with such orders by chance for small sample sizes due to huge sampling bias. We will have better chances to obtain reliable estimates if expert knowledge about the parameters can be incorporated into partially supervised learning.

- **Systematic comparison across different learning types** – We suggested AIC based comparison between the results from partially supervised learning methods and have shown that it fairly works. We however performed comparison of partially supervised learning to the others by either visualizing the results or conducting explanatory data analysis. From our simulations we found semi-supervised learning still best performs on some data sets even though it uses less information than partially supervised learning. Comparison of estimates based on visualization or explanatory data analysis requires much additional efforts. Moreover its performance depends on a person who analyzes data, so it is not reproducible. A method for systematic comparison of estimates from partially supervised learning to those from conventional learning methods will help people efficiently select the best estimates across learning classes.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Almuallim, H., Akiba, Y., and Kaneda, S. (1995). On handling tree-structured attributes in decision tree learning. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 12–20.

Ambroise, C. and Govaert, G. (2000). EM algorithm for partially known labels. In *Proceedings of the 7th Conference of the International Federation of Classication Societies*, pages 161–166.

Côme, E., Oukhellou, L., Denux, T., and Aknin, P. (2008). Mixture model estimation with soft labels. In Dubois, D., Lubiano, M., Prade, H., Gil, M., Grzegorzewski, P., and Hryniewicz, O., editors, *Soft Methods for Handling Variability and Imprecision*, volume 48 of *Advances in Soft Computing*, pages 165–174. Springer Berlin / Heidelberg.

Côme, E., Oukhellou, L., Denx, T., and Aknin, P. (2009). Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334–348.

Cour, T., Sapp, B., Jordan, C., and Taskar, B. (2009). Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition 2009*, pages 919–926.

Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536.

Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8(8):947–961.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1996). Inference about defects in the presence of masking. *Technometrics*, 38(3):247–255.

Flehinger, B. J., Reiser, B., and Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika*, 85(1):151–164.

Gill, R. D., Van Der Laan, M. J., and Robins, J. M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. *State of the Art in Survival Analysis, Springer Lecture Notes in Statistics*, 123:255–294.

Guess, F. M., Usher, J. S., and Hodgson, T. J. (1991). Estimating system and component reliabilities under partial information on cause of failure. *Journal of Statistical Planning and Inference*, 29(1):75–85.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8(3):431–444.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328):1459–1471.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, pages 2244–2253.

Hosmer, Jr., D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29(4):761–770.

Jewell, N. P. (1982). Mixture of exponential distributions. *The Annals of Statistics*, 10(2):479–484.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kuss, H. M., Lüngen, S., Müller, G., and Thurmann, U. (2002). Comparison of spark OES methods for analysis of inclusions in iron base matters. *Analytical and Bioanalytical Chemistry*, 374(7):1242–1249.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3):201–211.

Lin, D. K. J. and Guess, F. M. (1994). System life data analysis with dependent partial knowledge on the exact cause of system failure. *Microelectronics Reliability*, 34(3):535–544.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, volume 299. Wiley-Interscience.

Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, 45(3-4):504–520.

Miller, D. J. and Browning, J. (2003). A mixture model and em-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1468–1483.

Miyakawa, M. (1984). Analysis of incomplete data in competing risks model. *IEEE Transactions on Reliability*, 33(4):293–296.

National Cancer Institute (2011). Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2008), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2011, based on the November 2010 submission.

Papadapoulos, A. S. and Padgett, W. J. (1986). On Bayes estimation for mixtures of two exponential-life-distributions from right-censored samples. *IEEE Transactions on Reliability*, 35(1):102–105.

Park, C. (2005). Parameter estimation of incomplete data in competing risks using the EM algorithm. *IEEE Transactions on Reliability*, 54(2):282–290.

Peng, Y. and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243.

Ramon, J., Albert, G., and Baxter, L. A. (1995). Applications of the EM algorithm to the analysis of life length data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(3):323–341.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

Rider, P. R. (1961). The method of moments applied to a mixture of two exponential distributions. *The Annals of Mathematical Statistics*, 32(1):143–147.

Schlattmann, P. (2008). *Medical Applications of Finite Mixture Models*. Springer.

Sharma, R. and Poole, D. (2005). Probabilistic reasoning with hierarchically structured variables. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 1391–1397.

Szczurek, E., Biecek, P., Tiuryn, J., and Vingron, M. (2010). Introducing knowledge into differential expression analysis. *Journal of Computational Biology*, 17(8):953–967.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Usher, J. and Hodgson, T. (1988). Maximum likelihood analysis of component reliability using masked system life-test data. *IEEE Transactions on Reliability*, 37(5):550–555.

Usher, J. S. and Guess, F. M. (1989). An iterative approach for estimating component reliability from masked system life data. *Quality and reliability engineering international*, 5(4):257–261.

Vannoorenberghe, P. and Denœux, T. (2002). Handling uncertain labels in multiclass problems using belief decision trees. In *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, volume 3, pages 1919–1926.

Vannoorenberghe, P. and Smets, P. (2005). Partially supervised learning by a credal EM approach. In *Proceedings of the 8th European conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 956–967.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5(3):329–350.

Zhang, J. (2005). *Learning ontology aware classifiers*. PhD thesis, Iowa State University.

Zhang, J., Kang, D. K., Silvescu, A., and Honavar, V. (2006). Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data. *Knowledge and Information Systems*, 9(2):157–179.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130.